Ninth Biennial Report

April 2007 – April 2009

Contents

Part I

Overview – The Institute

1 Overview

Summary

The present report is a survey over the scientific activities of the Max Planck Institute for Informatics for the period April 2007 – April 2009. We begin with an overview of the institute's mission and goals and briefly highlight recent developments. The following sections provide individual progress reports by the institute's research groups.

Mission and Strategic Goals

The past three decades have brought dramatic changes in the way we live and work. This phenomenon is widely characterized as the advent of the Information Society. It is fueled by the power of information technology to acquire, store, process and transmit data compactly, inexpensively and at greater speeds than ever before. Ten years ago, most digital content was textual. Today, graphical and audiovisual input and output devices are in widespread use and modern personal computers have multimedia capabilities. Ubiquitous sensing devices will further increase the global volume of digital data. The availability of digital content in a variety of different formats and modalities and the increasingly pervasive access to the Internet combine to make a host of information available to anyone, at anytime.

Given these trends, the challenge of informatics research is to organize, understand, and search digital data in a robust, efficient and intelligent manner, and to create dependable systems that support natural and intuitive user interaction.

The Max Planck Institute for Informatics faces these challenges by cutting-edge research in informatics with a focus on algorithms and their applications. Our research ranges from foundations (algorithms and complexity, automation of logic) to a variety of application domains (computer graphics, geometric computation, constraint solving, databases and information systems, machine learning, computational biology and computational genomics and epidemiology).

- On the fundamental research side, this involves first-class research on new algorithms.
- The algorithmic work for applications encompasses the integration of new algorithms into full-fledged systems and concrete application scenarios that are of high practical relevance. This involves the implementation of comprehensive software platforms and their experimental evaluation in application settings.
- We provide a stimulating environment for junior researchers that enables them to develop their own research programs and build up their own groups.

Most of the major advances in informatics have come through the combination of new theoretical insights and experimental validation. Our goal is to have impact through publications, software and people alike.

The overarching vision of the MPI for Informatics, the other informatics institutes on campus (MPI for Software Systems, German Center for Artificial Intelligence DFKI), and the Department of Computer Science at Saarland University is to firmly establish the region as one of the top locations for research in informatics world-wide.

We are strongly committed to communicating our results. Exposing and testing our ideas to the research and development communities leads to improved understanding. We actively seek publication of our results and findings in professional journals and conferences, and we use our Internet pages to make our results widely available to the community. We seek users for our prototype systems among those with whom we have common interests, and we encourage collaboration with researchers from academia and industry alike. In selected domains we offer professional software and Internet services that are the basis of research and application by wide user communities.

We provide a stimulating environment for junior researchers that enables them to develop their own research programs and build up their own groups. The institute runs an active fellowship program on both the PhD and postdoc level and, since the establishment of the institute, a large number of researchers have spread out over other institutions, many of them taking tenured positions.

Situation and Recent Developments

There have been several important developments during the reporting period.

The institute is intimately involved in the *Cluster of Excellence on Multimodal Computing and Interaction* and the *Saarbrücken Graduate School of Computer Science* that were established at Saarland University within the framework of the German Excellence Initiative in 2007. The scope of the cluster has been strongly influenced by the research agenda of the institute, and all institute directors fully participate in both cluster and graduate school as principal investigators. The cluster of excellence is coordinated by Hans-Peter Seidel.

The cluster carries high risk, long term research plans and has two main goals: First, to enable natural multimodal interaction with information systems anytime and anywhere, exploiting the wealth of modalities present in everyday human-to-human interaction. Second, to enhance the ability of computer systems to acquire, process, and present different modes of data in an efficient and robust way. We aim for systems that can analyze and interpret multimodal information even when it is large, distributed, noisy and possibly incomplete; that can organize the obtained knowledge for powerful querying; and that can produce visual feedback in real time. We refer to this type of computing as multimodal computing. Total funding for the cluster of excellence and the graduate school amounts to about 40 M € over a five year period.

The institute is a founding member of the newly created *Intel Visual Computing Institute (IVCI)* that will be established on the campus of Saarland University. IVCI constitutes Intel's biggest university collaboration in Europe. The institute also remains a major partner in the *Center for Bioinformatics Saar (CBI)*, and two research groups of the institute actively

participate in the *DFG Transregio-SFB AVACS (Automatic Verification and Analysis of Complex Systems)* that has been renewed last year.

We have also further strengthened our international ties. In addition to the *Max Planck Center for Visual Computing and Communication* which establishes a stable link with Stanford University (see below), the Max Planck Society (MPG) has recently approved funding for several partner groups at India that will be established at IIT Delhi. Microsoft Research Cambridge has committed up to 15 PhD fellowships for our *International Max Planck Research School (IMPRS)*. Several major EU projects have been started (see below).

The leader of our independent research group on machine learning, Tobias Scheffer, has accepted an offer for a full professorship from Potsdam University and has left. Alice McHardy newly joined the institute as leader of an independent research group (IRG1) on Computational Genomics and Epidemiology in 2007.

Several senior researchers newly joined the institute. Most senior researchers chair their own group, and all senior researchers are members of the Extended Director's Board of the institute. Senior Researchers can be formally granted the right to independently supervise their own PhD students by the Faculty of Mathematics and Computer Science at Saarland University. At the time of writing, in addition to the directors, the senior researchers of the institute are Mario Albrecht, Holger Bast, Benjamin Doerr, Francisco Domingues, Elmar Eisemann, Khaled Elbassioni, Alice McHardy, Meinard Müller, Karol Myszkowski, Thomas Neumann, Ralf Schenkel, Viorica Sofronie-Stokkermans, Ingolf Sommer, Rob van Stee, Robert Strzodka, Martin Theobald, Thorsten Thormählen Michael Wand, and Christoph Weidenbach.

Presently 115 doctoral students and 60 postdocs are affiliated with the institute. The scientific staff is complemented by the administration with 16.75 members (including secretaries), the IST group (9 members of staff), our library (2 members of staff), and our technical support group (3 members of staff). The central administration group, the IST group and the technical support group is shared with MPI-SWS. The IST group operates various servers and clusters and a network of approximately 500 workstations and notebooks.

Teaching and Graduate Education

The institute makes a strong effort to offer a variety of courses to computer science and bioinformatics students of Saarland University. Courses taught during the period of this report are listed in Sections 28.14, 29.12, 30.14, 31.11, 32.11, and 34.8. Within the reporting period 48 doctoral dissertations and 4 habilitations have been completed successfully.

Teaching is partly embedded into the day-to-day offering of Saarland University and partly linked to special projects in undergraduate and graduate education which we will describe now.

The central stage of graduate research for the institute is the International Max Planck Research School for Computer Science (IMPRS-CS) which was founded in the year 2000. This is a graduate school in computer science (with courses taught in English) which offers both a Masters and a PhD program. About 50% of the PhD students come from foreign countries (main portions from Bulgaria, China, Greece, India, Romania and Russia). All funded Master students within the IMPRS program are foreigners. During the reporting period 2007-2009, 49 IMPRS-CS students have successfully completed their PhD.

Members of the institute together with their colleagues at the computer science department of Saarland University participate in the Graduiertenkolleg (Graduate Research Center) on Quality Guarantees for Computer Systems and the recently funded Saarbrücken Graduate School for Computer Science.

Since 2007 we receive additional funding for the IMPRS-CS including support from the German-French University for common PhD projects with Nancy (France) in the form of up to 10 mobility grants per year (each up to 18 months) and infrastructure support as well as a yearly support of up to 5 additional PhD scholarships from Microsoft Cambridge for the next 3 years.

Equal Opportunity Measures

We continue our efforts to increase the representation of women among our research staff. With Alice McHardy we won our first female independent research group leader.

In addition to the measures we have implemented now for several years, the "Girl's Day" and "University Camp for Pupils", and prioritizing female applications on all levels, we started the following projects.

We initiated a group of female researchers from our institute and have given them a budget in order to attract further women. The outcome of this group so far is an extra postdoc position for female researchers, a talk series of female researchers with an international standing (first talk by Claudia Perlich from IBM Research in June 2009) and an initiative to increase the number of childcare positions on campus (together with MPI-SWS).

In September 2008 we hosted the final round of the German computer science competition for pupils. All participants actually were boys. Our conclusions from this are twofold. Starting this year we will invite already the semi-finalists, having a reasonable portion of girls, together with their respective teachers to spend a computer science day on the campus. In addition, we issued an award for the best girl and best boy performance in computer science among the graduate classes of all high schools in the Saarbrücken region. Those pupils will be invited to our computer science day as well. Finally, in a joint effort together with our colleagues from the computer science department and the computer science students on campus, we have prepared slides for a talk to inspire pupils and in particular girls for computer science. The talk will be given by computer science students and researchers at their former high schools. Together with the computer science department we sponsor travel costs for the relevant speakers.

Third Party Funding

In addition to the strategic collaborations mentioned above, the institute participates in quite a number of projects funded by research grants awarded by the European Union (EU), the German-Israeli Foundation (GIF), the German Academic Exchange Service (DAAD), the German Research Foundation (DFG), and the German Ministry for Education and Research (BMBF), among others. Cooperation with industry has also substantially increased. For the descriptions of these grants see Sections 28.16, 29.14, 30.16, 31.13, and 32.13.

Max Planck Center for Visual Computing and Communication

The Max Planck Center for Visual Computing and Communication (MPC-VCC) with corresponding research activities at the MPI for Informatics and at Stanford University was established jointly by MPG and Stanford in 2003. The collaboration has two intertwined goals: Establish a joint research program in the area of visual computing and communication, and, incorporate a strong career development component to alleviate the shortage of qualified faculty and scientists in information technology in Germany.

Honors and Awards

Several outstanding honors and awards were conferred on members of the institute. Kurt Mehlhorn was awarded an honorary doctoral degree from the Faculty of Science at the Aarhus University, Denmark in 2008. Thomas Lengauer was named Member of the German Academy of Science and Engineering - acatech in 2007, Gerhard Weikum was named a Member of acatech in 2008.

Three scientists of the institute were conferred with the Otto-Hahn medal of the Max Planck society honoring outstanding dissertations: Martin Theobald (D5)and Christian Theobalt (D4) in 2007 and Sebastian Michel (D5) in 2008. Christoph Bock (D3) will receive the medal in 2009. Ivo Ihrke received the Eduard Martin Prize 2007 (PhD award by Saarland University). Martin Theobald (D5) also received the ACM SIGMOD Doctoral Dissertation Award Honorable Mention 2006 presented in 2007. Sebastian Michel (D5) was awarded with the GI/DBIS Dissertation Award 2007/2008.

Holger Bast received the Dr. Meyer-Struckmann Award 2007. The Heinz-Billing prize 2007, an award given bi-annually by the Heinz-Billing association for scientific computing inside the Max Planck society, was won by Holger Bast (D1) and Stefan Funke (D1) with their contribution "Ultrafast Shortest-Path Queries via Transit Nodes". They also received the Alcatel-Lucent Award and the Saar LB Science Award in 2008. Rolf Harren (D1) received the Hans-Uhde-Award 2008. Also in 2008 the ETH Medal for PhD Thesis was given to Konstantinos Panagiotou (D1). Marc Spaniol (D5) received the Friedrich-Wilhelm Dissertation Award in 2008. A Google Research Award 2008 was won by Tobias Scheffer (RG2). Hendrik Lensch received an Emmy-Noether-Fellowship from the German Research Foundation, and Bodo Rosenhahn won the DAGM Olympus Prize 2007. Michael Goesele (now with TU Darmstadt) and Christian Theobalt (now with Stanford) both received the Eurographics Young Researcher Award.

Paper and poster awards are mentioned specifically for each group in Part III of this report.

Cooperations

The cluster of excellence "Multimodal Computing and Interaction" is the major project at Saarbrücken in which all directors of the institute, a good portion of our senior researchers, Peter Druschel from our sister institute for software systems and the colleagues Matthew Crocker, Manfred Pinkal, Raimund Seidel, Philipp Slusallek, Hans Uszkoreit, Wolfgang

Wahlster, and Joachim Weickert from Saarland University are principal investigators and cooperate in seven different research areas.

Inside the institute and in addition to the cluster activities, there are the following cooperations.

- D1 and RG1 jointly participate in AVACS. Together they investigate the integration of LP solving technology into hierarchic theorem proving.
- D3 implements with D4 molecular dynamics algorithms on graphics hardware.
- D3 cooperates with D1 on parsimonious modeling of protein networks.
- D3 cooperates with RG2 on the topics of developing methods for bias-aware learning in HIV therapy outcome prediction.
- D4 and D1 cooperate on surface reconstruction and shape matching.
- D5 cooperates with D1 on efficient algorithms for information retrieval.
- D5 cooperates with RG2 on machine learning for text and Web mining.
- RG1 cooperates with D5 on complete algorithms for reasoning about ontologies.
- RG1 cooperates with Andrey Rybalchenko from MPI-SWS on abstraction refinement methods.
- IRG1 cooperates with D3 on the computational study of viral disease and evolution.
- IRG1 cooperates with RG2 in adapting machine learning techniques to problems in computational genomics.

All groups have strong relations to researchers at Saarland university.

- D1 and RG1 cooperate with Bernd Finkbeiner, Holger Hermanns, and Reinhard Wilhelm from Saarland University in the transregional collaborative research center AVACS.
- D1 cooperates with the group of Raimund Seidel on algorithms for geometric problems.
- D1 cooperates with the group of Markus Bläser on bio-inspired algorithmic methods.
- D3 cooperates with the bioinformatics groups of Saarland university in the Center for Bioinformatics (CBI).
- D4 cooperates in the areas of realtime ray tracing and interactive global illumination (Ph. Slusallek), mesh processing and 3D video processing (J. Weickert), speaker localisation (T. Herfet), speech synchronization (M. Pinkal, H. Uszkoreit), and modeling and animation of synthetic virtual characters (W. Wahlster)
- D5 cooperates with Hans-Peter Lenhof on scalable middleware for biological network analysis.

Professional Activities

Members of the institute were involved in the organization of 22 workshops and conferences. In 190 cases we have been invited to join the program committee of major international conferences, not counting program committee memberships for national conferences and international workshops. Finally, we serve on the editorial boards of 25 scientific journals.

Public Outreach

Since its foundation the Max Planck Institute for Informatics has reached a considerable level of awareness in the national and international research community. In order to improve the institute's visibility also to the general public and in particular to young academics, several measures have been implemented in the recent two years.

In 2009 our workshops, see our equal opportunity measures on page 6, will be extended by the "3sat Nano Camp". Accompanied by a camera crew 12 pupils attend four different workshops hosted by the institute for one day and experience computer science research through several experimental setups. The resulting movie will be broadcasted by the national TV stations 3sat and ZDF.

The institute exhibited innovative software products on the CeBIT fair 2008 and 2009 in Hannover and participated in several cultural exhibitions both, nationwide ("Science Ship" 2009) and regional ("Science Summer" 2009, Open House 2008). Specifically, these public exhibitions focus on attracting diverse and varying audiences. Besides regular press releases and continuously driven presence in local and nationwide newspapers (Campus Extra 2008, " TRENDS Magazine Saar" 2008, Max-Planck-Research 2008) the institute specifically aims to strengthen the collaboration with high level and technically specialized journalists. Among other measures a competition for computer science journalists (Computer Science Journalist Award 2008 and 2009) has been implemented in close collaboration with the computer science department, our sister institute for software systems and the DFKI. But also groups of national and international journalists were invited on a regular basis for several days to attend presentations and workshops on cutting-edge research topics.

At a glance, all recent measures attached importance on raising the institute's level of awareness within different target groups. Through various channels we reached distinct audiences, such as the general public, pupils and young academics, and also professional computer science journalists. In the future we plan to intensify this strategy.

Offers of Faculty Positions

The following members of the groups received offers for faculty positions or equivalent positions within the reporting period:

Ben Adams (IRG1, University of Bath)
Mario Albrecht (D3, University of Kiel)
Ernst Althaus (D1, University of Mainz)
Omid Amini (D1, École Normale Supérieure, Paris)
Iris Antes (D3, University of Bergen (declined), University of Munich)
Hannah Bast (D1, University of Mainz, University of Freiburg)
Alexander Belyaev (D4, Heriot-Watt University Edinburgh)
Benjamin Doerr (D1, RWTH Aachen University (declined), University of Dortmund (declined))
Hongbo Fu (D4, City University of Hong Kong)
Stefan Funke (D1, University of Greifswald)
Joachim Giesen (D1, University of Bonn (declined), University of Jena)

Michael Goesele (D4, University of Darmstadt)
Thorsten Grosch (D4, University of Magdeburg)
Mouna Kacimi (D5, University of Bolzano)
Hendrik Lensch (D4, University of Ulm)
Bodo Rosenhahn (D4, University of Hannover)
Tobias Scheffer (RG2, University of Freiburg (declined), University of Chemnitz (declined), University of Potsdam)
Viorica Sofronie-Stokkermans (RG1, University of Manchester (declined))
Christos Tryfonopoulos (D5 , University of Peloponnese)
Rhaleb Zayer (D4, INRIA Lorraine, France)

Part II

Overview – The Research Units & Senior Researchers

2 The Algorithms and Complexity Group (D1)

History

The algorithms and complexity group (D1) was established in 1990 as one of the two founding groups of the institute. Kurt Mehlhorn leads the group since its foundation.

Group Composition and Development

The senior scientists and subgroup coordinators are Hannah Bast, Benjamin Doerr, Khaled Elbassioni, Kurt Mehlhorn, Frank Neumann, Michael Sagraloff, and Rob van Stee. Kurt Mehlhorn and Benjamin Doerr are on permanent contracts, the others on five year contracts. During the report period, Ernst Althaus, Stefan Funke and Jochim Giessen moved to tenured professorships in Mainz, Greifswald and Jena, respectively. Hannah Bast will move to Freiburg, while Rob van Stee joined the group from Karlsruhe. The senior scientists and subgroup coordinators set the scientific direction of the group, and select the PhD-students and postdocs.

There are currently 17 researchers and 14 PhD students in the group. Researchers are typically on two year contracts. We continue to attract excellent postdocs in our annual application round. 16 PhD students graduated in the reporting period. We will make an effort to attract additional PhD and master students. Section 28.1 lists the names of current and recent group members and the current positions of the group members that left during the report period.

Vision and Research Strategy

The overarching vision of the Max Planck Institute for Informatics, the other informatics institutes on campus (MPI for Software Systems and German Center for Artificial Intelligence), and the Fachbereich Informatik of Saarland University is to turn the region into one of the top ten locations for research in informatics world-wide. The specific vision for D1 is:

- Be known among the peers as one of the first class algorithm groups and as a trend setter in parts of algorithmics.
- Have impact on the research community and society through research results, people, software, and scientific leadership.
- Address current challenges in algorithmics.

The following challenges motivate much of our work.

- Understand randomness and quasi-randomness.
- Understand the structure of hard combinatorial problems.
- Develop a theory of bio-inspired computing.
- Show that reliable and efficient geometric computing in the 3d-domain is feasible.
- Develop intelligent and efficient search techniques for large data sets.
- Develop methods for decision making under uncertainty (on-line, insufficient computational resources, partially unkwown input, selfish agents).
- Develop efficient algorithms for central combinatorial and geometric problems.

About two-thirds of our effort is theoretical work. One-third is experimental and software construction. We believe that implementation is an essential part of algorithmic research for two reasons: first, experiment brings out hidden assumptions and poses new questions, and, second, implementations strongly increase the impact of theoretical work. We pursue some of our implementations all the way to systems that are useful to others. For instance, CompleteSearch serves the community as a search engine in DBLP, the computational geometry algorithms library CGAL is widely used for geometric computations, and LEDA, the library of efficient algorithms and data structures, is still downloaded more than a dozen times per day.

The research strategy is set partially top-down and partially bottom-up. Some research lines are defined by the senior scientists. Others are defined by the individuals in the group. Particularly successful themes of individuals may turn into themes for the group. Current examples are bio-inspired computing (Frank Neumann and Benjamin Doerr) and information retrieval (Hannah Bast). Past examples are graph drawing, algorithm engineering, bioinformatics, and external memory computation.

There are three lines of research that have been part of our portfolio since the creation of the group: foundations, computational geometry and geometric computing, and combinatorial algorithms and software libraries. We make sure that these lines of research are always sufficiently represented in the group. However, the exact nature of the research under these headings has changed over time.

In the foundations area, our current emphasis is on understanding randomness and quasi-randomness and in understanding the structure of hard combinatorial problems. The foundational work on bio-inspired computing is closely linked to understanding randomness.

With respect to computational geometry and geometric computing, the emphasis has shifted towards geometric computing, i.e., the science of reliable and efficient geometric software. Since about 2002, we are making a major and coordinated effort in extending our past work on linear geometry to non-linear geometry. We perform part of the work in international collaborations (CGAL and a GIF project with Tel Aviv University).

With respect to combinatorial algorithms, the emphasis is currently on theoretical work on polyhedra, decision making under uncertainty, and algorithmic game theory and applied

work on efficient control of OLED displays, problems in bioinformatics, and exact linear and integer programming.

Examples of scientific leadership are our involvement in the cluster of excellence on *Multimodal Computing and Interaction* and the establishment of the DFG (= German Research Foundation) priority program *Algorithm Engineering*. Ernst Althaus, Hannah Bast, and Benjamin Doerr are funded under this program.

Research Areas and Achievements

We will report about our research under the headings

- Foundations and Discrete Mathematics (Coordinator: Benjamin Doerr)
- Bio-inspired Computation (Coordinator: Frank Neumann)
- Computational Geometry (Coordinators: Stefan Funke and Joachim Giesen)
- Geometric Computing (Coordinator: Michael Sagraloff)
- Combinatorial Optimization (Coordinators: Khaled Elbassioni and Kurt Mehlhorn)
- Algorithmic Game Theory and Online Algorithms (Coordinator: Rob van Stee)
- Information Retrieval (Coordinator: Hannah Bast)

However, *we are not organized into disjoint subgroups and there is little hierarchy.* Many of us contribute to several areas. Also, the group coordinators lead by virtue of scientific quality and leadership and not by virtue of their position in the hierarchy.

We will next highlight our most exciting results of the past two years. We have published not only in algorithms conferences and journals, but also in information retrieval and artificial intelligence.

Kernelizations for Generic Optimization Problems (STACS 2009)

Investigator: Stefan Kratsch

We provide efficient kernelizations, the formal notion of preprocessing from parameterized complexity, for two classes of syntactically defined optimization problems, both known to be constant-factor approximable (APX). Our work strengthens the perceived relation between kernelization and constant-factor approximability. It is open to extend polynomial kernelization bounds to larger subclasses of APX. See Section 28.4.2.

Analysis of Random Planar Graphs (SODA 2009, full version accepted for TALG):

Investigator: Kosta Panagiotou

Let C be a class of labeled connected graphs, and denote by B_n the number of vertices in a largest block (i.e. maximal biconnected subgraph) of a random graph from C that contains exactly n vertices. We discover, depending on some critical condition of the generating function for C, that B_n contains with high probability either a constant fraction of the vertices, or is at most of logarithmic size. In particular, random planar graphs are of the first kind and random outerplanar graphs are of the second kind. See Section 28.4.3.

Quasirandom Broadcast Protocols (SODA 2008, ALENEX 2009, ICALP 2009)

Investigators: Benjamin Doerr, Tobias Friedrich

In these works, we develop and analyze a quasirandom protocol for broadcasting a piece of information to all nodes of a network. Mathematical proofs and experimental evaluations show that it has a number of advantages over the classical solutions to this problem. To the best of our knowledge, this is the first time that quasirandomness, which was very successful in numerical integration, is used in computer science. More details can be found in Sections 13 and 28.4.5.

Crossover is Provably Useful (GECCO 2008)

Investigators: Benjamin Doerr, Edda Happ, Christian Klein

In evolutionary computation, existing solution candidates (called individuals) are used to generate new ones via small local changes to one solution (mutation) and by combining two solutions to one (crossover). While experimental results leave little doubt that both principles are useful, so far there was no convincing proof for the fundamental question whether crossover can bring an advantage. We answer this long disputed problem by showing that for the classical all-pairs shortest path problem, the natural evolutionary algorithm has an asymptotically better run-time if it can use crossover. For more details, see Section 28.5.2.

Fixed-Parameter Evolutionary Algorithms (GECCO 2009)

Investigators: Stefan Kratsch and Frank Neumann

We considered evolutionary algorithms for the first time in the context of parameterized complexity. Investigating the vertex cover problem, we related the runtime of evolutionary algorithms to the input size and the cost of a minimum solution, and pointed out that the search process of evolutionary algorithms using multi-objective models creates partial solutions that are similar to the effect of a kernelization. Based on this, we showed that evolutionary algorithms are randomized fixed-parameter tractable algorithms for the vertex cover problem. See Section 28.5.2.

Small ϵ-nets (SoCG'08)

Investigators: Evangelia Pyrga and Saurabh Ray

We give a new method for showing the existence of small size epsilon nets which we also use to prove new results and to considerably simplify older ones. The method is based on characterizing geometric objects in a more combinatorial way. Several results can then be obtained using a very simple double counting argument. As a consequence, the constant involved drops from over a million to around 30, resulting in a corresponding improvement in approximation algorithms that depend on Epsilon-nets. See page 158.

Root Isolators (Theoretical Computer Science 2008, ISSAC 2009)

Investigators: Arno Eigenwillig, Kurt Mehlhorn, Michael Sagraloff, Vikram Sharma

We prove the first polynomial time bound on the bit complexity of the continued fraction approach to root isolation of univariate polynomials with integer coefficients. We also give the first deterministic root isolator for polynomials with bitstream coefficients. See Section 28.7.2.

Algorithms for Algebraic Curves and Surfaces (SoCG 2008, CGTA to appear, ESA 2007, SPM 2008)

Investigators: Eric Berberich, Michael Kerber, Kurt Mehlhorn Michael Sagraloff

We designed new algorithms and the corresponding software for constructing subdivisions induced by curves on parametric surfaces, such as spheres, tori, and cylinders. The subdivisions support point location and ray shooting queries. We also present a method for computing the exact topology of a real algebraic surface, together with the geometric information about critical points. We can also compute a simplicial complex, isotopic to the surface. See Sections 28.7.3 and 28.7.4 for more details.

A Lower Bound for Weighted Flow Time (SODA 2009)

Investigator: Ho-Leung Chan

We considered the classic online scheduling problem of minimizing the total weighted flow time on a single machine with preemptions. Here, each job j has an arbitrary arrival time r_j, weight w_j and size p_j, and given a schedule its flow time is defined as the duration of time since its arrival until it completes its service requirement. It was widely believed that the problem admits a constant factor competitive algorithm. We disproved this belief and showed the first $\omega(1)$ lower bound on the competitive ratio of any deterministic online algorithm, hence closing one of the most important open problem in online scheduling. Please see Section 28.9.4 for more details.

Approximating Two-Dimensional Bin Packing (submitted)

Investigators: Rolf Harren and Rob van Stee

Two-dimensional bin packing is a well-known geometric generalization of classical bin packing where a list of rectangles has to be packed into a minimal number of equally-sized rectangular bins. For the problem where we do not allow rotations of the items we show that a packing into at most twice the optimal number of bins can be found in polynomial time. As we can not get a better approximation unless P = NP, our result settles the question of absolute approximability of this problem. More details can be found in Section 28.8.1.

The Ultimate Analysis of LRU, (SODA 2009)

Investigators: Spyros Angelopoulos and Pascal Schweitzer

It is well-known that competitive analysis is not always consistent with the empirical evaluation of online algorithms. The most notable example is the paging problem, where there is a vast class of algorithms with the same competitive ratio, ranging from naive to sophisticated ones. We applied a powerful, yet simple and intuitive form of analysis termed *bijective analysis* for the paging and list update problems. Our main result is a theoretical justification of the superiority of the Least-Recently-Used strategy, see Section 28.9.4.

Adaptive Local Ratio (SODA 2008)

Investigator: Julián Mestre

Local Ratio is a well-known paradigm for designing approximation algorithms. A critical step in these algorithms is to decompose the input objective function into simpler easy-to-deal-with

objectives. This decomposition is normally done in a fixed, pre-specified manner. We proposed an *adaptive* variant of the Local Ratio technique, which turns the problem of finding a good decomposition into an optimization problem of its own. We showed how this technique can lead to provably better approximations for a data migration problem. See Section 28.8.1 for more details.

Checking Self-duality of Polytopes (SoCG 2008)

Investigators: Khaled Elbassioni, Hans Raj Tiwary

We study the problem of checking whether a polytope given by its vertices or facets is combinatorially isomorphic to its polar dual. We show that the problem is Graph Isomorphism hard, and that it is Graph Isomorphism complete if and only if vertex enumeration is Graph Isomorphism easy. We also proved that checking self-duality of a polytope, given by its facet-vertex incidence matrix, is Graph Isomorphism complete, see Section 28.8.4.

Mature Geometric Software (CGAL packages)

Investigators: Michael Hemmer, Eric Berberich, Michael Kerber, Pavel Emeliyanenko

We contributed five packages to CGAL, in particular, an algebraic kernel. The packages offer, for example, number types, polynomials, root finders, analysis of curves and surfaces, and arrangement computation for curves. Our contributions constitute a significant part of the current effort of extending CGAL from linear to non-linear objects, see Section 28.7.

CompleteSearch (deployed at DBLP)

Investigators: Hannah Bast, Ingmar Weber, Daniel Fischer, Markus Tetzlaff, Marijan Celikik

The CompleteSearch engine became an alternative search engine for DBLP, a popular site for literature search in CS. Try it at `http://dblp.mpi-inf.mpg.de/dblp-mirror/index.php`. It offers fast search (a search is triggered after each keystroke, with instant response times), prefix search (sig matches SIGIR), error-tolerant search (probabilistic will also match the misspelling "probalistic"), faceted search (at any time, the right sidebar gives a breakdown of the current result set (initially: all entries) by various categories; click an item to refine), exact word match (end word with a dollar, e.g. graph$), boolean or (put a pipe between words, e.g., graph—network), and phrase search (put a dot between words, e.g., information.retrieval). See Section 28.10.

Algorithms for Longer OLED Lifetime (turned into a chip)

Investigator: Andreas Karrenbauer

We designed and analyzed new controllers for passive matrix OLED displays. Our approach is based on the fact that the degradation of such devices may be reduced by a suitable decomposition of the displayed images. We provide an approximation algorithm that was incorporated into a chip fabricated by Dialog Semiconductors to drive a 3 inch W-QVGA OLED display of TDK. See page 198.

Projects and Cooperations

The group is part of the excellence cluster "Multimodal Computing and Interaction".

We have published with departments D4 and D5 of the institute and the groups of Raimund Seidel and Markus Bläser in the Informatics Department of Saarland University.

Cooperations with Uri Zwick on graph algorithms and Dan Halperin on geometric computing are sponsored by GIF grants (German-Israel Foundation), cooperations with Naveen Garg (IIT Delhi) and Telikepalli Kavitha (IISc, Bangalore) under the MPG partner group program. Our work on geometric computing is embedded into the CGAL-project. The ACS project ended in 2008.

Ernst Althaus and Kurt Mehlhorn are part of the collaborative research center AVACS (Automatic Verification and Analysis of Complex Systems) funded by the German Research Foundation (DFG).

Prizes and Awards

Hannah Bast and Stefan Funke received the Heinz Billing Award 2007 and the SaarLB Wissenschaftspreis 2008 for their work on efficient route determination in street networks and Hannah Bast received the Meyer-Struckmann-Science Award 2008 and the Alcatel-Lucent Award 2008 for Technical Communication for her work on efficient intelligent search.

We received a number of best paper awards. Amr Elmasry and Julián Mestre are recipients of Humboldt fellowships, Rolf Harren and Kosta Panagiotou received prizes for their master- and PhD-thesis, respectively, and Kurt Mehlhorn received an honorary doctorate degree from Aarhus University. See Section 28.15.4.

Personal Remarks

Kurt Mehlhorn served as vice president of the Max Planck Society from 2002 to 2008. My term as vice president ended in June 2008. The vice presidency required about 25 to 30 working hours per week and implied frequent traveling. I was out of town for at least two days per week. My research time and research output was reduced accordingly. In particular, I did not read enough during this period. Also, my contact to students was reduced because I did almost no teaching. I want to thank the staff of the Algorithms and Complexity group for keeping the research in the department at a very high level and my co-directors for relieving me from administrative work in the institute.

My best works during the reporting period are

- the book "Data Structures and Algorithms: The Basic Toolbox" with Peter Sanders
- the deterministic Bitstream root isolator (see Section 28.7.2) and

– the very recent $O(m^\omega)$ algorithm for minimum cycle bases in graphs (see Section 28.8.3).

Since July 2008, I am again a full time member of the institute[1]. The figure shows the banner with which the institute welcomed me back. I learned a lot during my vice presidency, I took part in important decisions, but I am happy to be once more a full time researcher. My personal research is going well[2], and I am again fully integrated into the group, the institute, and the department.

My personal goals for the department for the next two years are: continue my research work on geometric computing and combinatorial optimization[3], establish more research collaborations (within and outside the institute), attract a larger number of PhD-students, and strengthen our visitor program.

[1] Of course, I am still serving on a number of external committees. In particular, I am a member of the Senate of the Max Planck Society and, since November 2008, I serve as interim head of the Max Planck Digital Library. These extra commitments take less than one day per week.

[2] I have finished papers on root isolation, cycle bases, algorithms for referee assignment, and graphlet counting.

[3] In particular, I want to address the following questions in the near future:

- Explain the behavior of the continued fraction method for root isolation. In practice, it performs much better than bisection, however, the proven bounds are weaker.
- Advance reliable geometric computing to the full 3d-domain.
- Fully exploit the power of controlled perturbation.
- Complete a book on Efficient and Reliable Geometric Computing with Chee Yap.
- Complete a survey article on certifying algorithms.
- Integrate our theoretical work on referee assignment into EasyChair.
- Find an $\tilde{O}(nm)$ algorithm for minimum cycle bases.

3 The Computational Biology and Applied Algorithmics Group (D3)

History

The Department became established when Thomas Lengauer joined the institute on October 1, 2001. During the reporting period, the Department has been essentially stable in size but experienced significant turnover. As this text is written, the Department has 21 scientists (5 postdoc, 16 predoc), two predoc and two postdoc fellowship holders, two support people and one guest, see `http://www.mpi-inf.mpg.de/departments/d3/people.html`. Two postdocs have part-time positions. Ten people have left and nine have joined the Department during the reporting period, respectively. Due to the fact that the curricula on bioinformatics at Saarland University have reached their steady state, there is a continuous influx of students into the Department who work on bachelor, diploma and master theses. Aside from the director, several scientists have offered in-depth courses on bioinformatics themes within the bioinformatics curricula at Saarland University.

Research Areas and Achievements

In the reporting period, the Department has basically continued along the themes and vision that it has established in the previous years. The main driving theme remains Computational Biology as it applies to diagnosis, prognosis and therapy of human disease.

Work on **HIV/AIDS** (see Section 29.4) has gained additional momentum through the successful completion of the EuResist project, which could facilitate a resistance database that spans several European countries and is significantly larger than what was available to us before. On the method side, the EuResist project has brought forth a new server that combines predictions by three tools [14], one being our tool THEO [1, 2]. The EuResist project was declared project of the month in the eHealth Monthly Focus of the EU in November 2008. Funding for the EuResist project has run out but the consortium has persisted. EuResist has incorporated and is partner in the new large EU-funded integrated project CHAIN. Furthermore, the consortium now receives some industry funding and is developing visions for a future that is independent of particular funding actions.

Thematically the analysis of viral coreceptor usage has caught increased attention both in the Department and in the medical community, in general [12]. The main reason is that this analysis is helpful or even required when administering new drugs targeting viral cell entry (see Section 29.4.2). Presently our internet server GENO2PHENO[coreceptor] is the dominant method for predicting viral coreceptor usage from viral genotype and the only one that we know to be used in clinical routine [17]. The hit numbers of 70,000 in the last two years bear support for this assessment.

A second thematical point is to enhance the existing statistical models for analyzing viral evolution in response to drug therapy be complemented with mechanistic models that implement concrete biological insight into the virus-host interactions. We have started a project towards such a goal with German cooperation partners

Another perspective in the field of HIV on which we are currently working is to bring our server up to processing genotypic data based on ultra-deep sequencing (454) technology.

On the distribution side, our HIV project will gain substantial additional visibility shortly. The project was the topic of the Public Lecture of the Max Planck Annual Meeting in Dresden in June 2008. Furthermore, several review articles have been solicited by both German and international journals with sizeable readerships (GI Informatik Spektrum, Spektrum der Wissenschaft, Communications of the ACM).

Our research on the **Hepatitis C virus (HCV)** (Section 29.5.2) focuses on two themes. On the one hand, we are currently preparing a database and tool pipeline in the style of GENO2PHENO to predict viral resistance against anti-HCV drugs from clinical resistance data. On the other hand, we continue developing and applying bioinformatics methods to elucidate the structure and function of HCV proteins and virus-host interactions. One point in case is a structural study of the resistance development of particular HCV variants against the new HCV-protease inhibitor telaprevir (Section 29.7.4) [19].

Computational epigenetics (Section 29.6) has thrived during the report period with several important publications on inferring epigenetic states of CpG islands from DNA sequence [6], constructing an epigenetic cancer biomarker [13], and analyzing inter-individual DNA methylation [7]. We also published a well cited review paper on computational epigenetics [5]. Christoph Bock who has pioneered this topic in the Department even as a doctoral student has been nominated group leader in the Department after his graduation in 2008. He since moved to Harvard University and Broad Institute but is still co-directing this group part-time. On the software side the milestone in this field is the complex statistical genome and epigenome browser EPIGRAPH which affords the advanced analysis of genome-wide datasets [4].

Additional topics of the Department include molecular networks (Section 29.5), the analysis of protein function from protein structure (Section 29.7) [18], molecular docking, with focus on protein flexibility [10] and on immuno-informatics (Section 29.8), as well as the analysis of genomics and transcriptomics data, partly with a focus of diagnosing types of cancer (Section 29.9) [16, 11]. These fields are continuing to draw solid contributions and reflect the global progress in their respective fields.

The Department and the Center of Bioinformatics Saar

The Department is a major partner in the Center of Bioinformatics Saar (CBI, `http://www.cbi-saar.de`), a joint initiative of Saarland University, the Max Planck Institute for Informatics and the Fraunhofer Institute for Biomedical Engineering, Sankt Ingbert, which has received substantial funding from DFG. DFG funding has now run out but the Center continues to operate with on-board resources and third-party funding of the individual partners. One reason that this is possible is Saarland University's continued strong commitment to Bioinformatics.

CBI Saar encompasses the complete structure of a bioinformatics center covering research

and teaching. Research is facilitated through a number of cooperative projects between biologists/chemists/medics and computer scientists. Teaching encompasses full curricula towards the bachelor and master of bioinformatics. Thomas Lengauer is continuing to be the scientific director of the Center. The Department engages in many of the Center's research projects and also provides much of the teaching.

Outreach and Service

In addition to its scientific contributions the Department has carried out several outreach and service projects. Thomas Lengauer was the Chair of the ISMB/ECCB 2007 Conference which took place in Vienna in July 2007. This meeting drew about 2000 delegates, had over 200 oral presentations and about 1000 posters. It was the largest conference on computational biology during the year.

On the topic of computational analysis of HIV resistance, Thomas Lengauer is one of the organizers of the annual Arevir Workshop, which takes place every April in Bonn, draws about 100 delegates and has turned into an international forum serving a dual purpose. On the one hand, the meeting is a certified continued-education event for German doctors. On the other it has become a meeting point for European researchers on HIV resistance and beyond (moving into HCV and HBV).

Thomas Lengauer is also preparing the Annual Meeting of the German Academy of Sciences Leopoldina, which takes place in October 2009 in Halle and has the theme "Computer Models in the Sciences - Between Analysis, Prediction and Suggestiveness". It will feature 15 keynotes from a wide variety of disciplines. The preliminary program has been determined. The meeting usually draws upwards of 700 participants.

The Department is hosting a number of information, registration and networking offers on the GISAID Platform (Global Initiative for Sharing Avian Influenza Data, `www.gisaid.org`). This is an independent multinational initiative that offers sequence and tracking data in influenza strains freely over the Internet. The EpiFlu sequence database of this initiative is hosted by the Swiss Institute for Bioinformatics. The GISAID site has been live since June 2008. The EpiFlu database has been used to select the strains for producing influenza vaccines for the Southern hemisphere in September 2008 and for the Northern hemisphere in February 2009.

Projects and Cooperations

Collaboration and Networking

There is ongoing cooperation with the other departments in the Institute on the topics of (i) molecular dynamics algorithms on graphics hardware (together with the Computer Graphics Department) and (ii) parsimonious modeling of protein networks (together with the Algorithms and Complexity Department). Furthermore, we have an active cooperation with the Machine Learning Group on the topic of developing methods for bias-aware learning in HIV therapy outcome prediction. This cooperation was established shortly after the arrival of that group at the Institute. Even though the Machine Learning Group has since left the Institute, the cooperation with them is persisting.

For a lab in Computational Biology, being placed within a computer science context requires special efforts of networking to the biology community. On campus the major device for networking to biologists is CBI. Our work on epigenetics and chemical biology, is in cooperation with CBI partners. The DFG Clinical Research Group on HCV used to be our major vehicle for cooperating with medical experts at Saarland University. Now these partners have moved to the Johann-Wolfgang-Goethe University in Frankfurt, but the cooperation persists and continues to be funded by DFG. On a national scale the GenaFor consortium on computational analysis of HIV resistance is continuing to cohere beyond the initial funding period. The same holds for the EuResist consortium on a European scale.

Our work on the analysis of structure-function relationships in proteins has been largely carried out within the European BioSapiens Network of Excellence whose funding will terminate in the middle of 2009.

New projects which have been acquired during the report period include the EU Integrated Project CHAIN (HIV resistance analysis), the EU project CancerDIP (Epigenetic analysis of cancer), and the BMBF Project HIV Cell Entry (Experimental and computational analysis of HIV cell entry).

Academic cooperation partners that are external to the institute are listed in the following. The list starts with cooperation partners of the group *Molecular Networks in Medical Bioinformatics*, followed by partners of the group *Computational Epigenetics* and a list of further collaborations of the Department.

- Dr. Melissa Cline (Center for Biomolecular Science and Engineering, University of California Santa Cruz, CA, USA) – Alternative splicing and protein networks
- Prof. Dr. Juana Díez (Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain) – Neurodegenerative diseases and protein function
- Dr. Robert Finn (Wellcome Trust Sanger Institute, Hinxton, UK) – DASMI protocol and protein domain interactions (EC-funded BioSapiens network)
- Dr. Henning Hermjakob (European Bioinformatics Institute (EBI), Hinxton, UK) – DASMI protocol and protein interactions (EC-funded BioSapiens network)
- Prof. Dr. Reinhard Jahn (Max Planck Institute for Biophysical Chemistry, Göttingen) – Neuronal endocytosis and protein function and interaction
- Dr. Silvia Krobitsch (Max Planck Institute for Molecular Genetics, Berlin) – Neurodegenerative diseases and protein function and interaction
- Prof. Dr. Young-Ae Lee (Max Delbrück Center for Molecular Medicine, Berlin) – Autoinflammatory diseases and protein structure-function (BMBF-funded NGFN)
- Prof. Dr. Petra Mutzel (Department of Computer Science, Dortmund University of Technology) – Graph layout algorithms for biological networks
- Dr. Andreas Prlić (Wellcome Trust Sanger Institute, Hinxton, UK) – DASMI protocol (EC-funded BioSapiens network)
- Dr. Markus Ralser (Max Planck Institute for Molecular Genetics, Berlin) – Neurodegenerative diseases and protein function and interaction

- Prof. Dr. Stefan Schreiber (Institute for Clinical Molecular Biology, University of Kiel) – Autoinflammatory diseases and protein structure-function (BMBF-funded NGFN)
- Dr. Frank Takken (Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands) – Plant disease resistance and protein structure-function
- Prof. Dr. Silvio Tosatto (Department of Biology, CRIBI Biotechnology Center, University of Padova, Italy) – Prediction of protein structure and function
- Prof. Dr. Alfonso Valencia (National Cancer Research Centre, Madrid, Spain) – Alternative splicing and protein networks (EC-funded BioSapiens network)
- Prof. Dr. Stefan Zeuzem (Faculty of Medicine, University of Frankfurt am Main) – Hepatitis C and viral proteins (DFG-funded Clinical Research Group)

- Prof. Esteller (Bellvitge Institute for Biomedical Research, Barcelona, Spain) – Cancer epigenetics
- Dr. Fuks (Free University of Brussels, Brussels, Belgium) – Cancer epigenetics
- Prof. Meissner (Broad Institute of MIT and Harvard) – Epigenome analysis
- Dr. Mikeska (Peter MacCallum Cancer Centre, Melbourne, Australia) – Cancer epigenetics
- Prof. Nekrutenko (Penn State University, University Park, PA, USA) – Epigenome analysis
- Dr. Waha (Department of Neuropathology, University of Bonn, Germany) – Cancer epigenetics
- Prof. Walter (Genetics/Epigenetics, Saarland University, Germany) – Epigenome analysis

- Prof. Beerenwinkel (ETH Zürich) – Bioinformatics for HIV
- Prof. Bernhardt (Biochemistry, Saarland University) – Docking and Homology Modeling
- Prof. Berthold (Pediatric Oncology, Cologne University Clinic) – Genomic Aberrations in Cancer
- Peter Bogner (Santa Monica) – GISAID
- Dr. Cox (US Centers of Disease Control and Prevention, Atlanta) – GISAID
- Martin Däumer (Laboratory Dr. Thiele, Kaiserslautern) – Bioinformatics for HIV
- Dr. Erhardt (University Hospital Düsseldorf) – Bioinformatics for HIV
- Dr. Fessel (Medical Care Program Northern California) – Bioinformatics for HIV
- Prof. Giffhorn (Applied Microbiology, Saarland University) – Docking and Homology Modeling
- Dr. Harrigan (British Columbia Center for Excellence in HIV/AIDS, Vancouver) – Bioinformatics for HIV

- Prof. Hartmann (Pharmacology, Saarland University) – Docking and Homology Modeling
- Dr. Heckmann-Pohl (Applied Microbiology, Saarland University) – Docking and Homology Modeling
- Dr. Kaiser (Virology, University of Cologne) – Bioinformatics for HIV
- Dr. Klau (Centrum Wiskunde & Informatica, Amsterdam) – Protein Structure and Function
- Dr. Krackhardt (Institute of Molecular Immunology, German Research Center for Environmental Health, Munich) – Docking and Homology Modeling
- Prof. Kräusslich (Virology, University of Heidelberg) – HIV Cell Entry
- Prof. Lackner (Molecular Biology, University of Salzburg) – Protein Structure and Function
- Prof. Lenhof (Bioinformatics, Saarland University) – Bioinformatics Software and Teaching
- Prof. Münk (Clinic for Gastroenterology, Hepatology and Infectiology University Hospital Düsseldorf) – Bioinformatics for HIV
- Prof. Pfister (Virology, University of Cologne) – Bioinformatics for HIV
- Prof. Rahnenführer (Department of Statistics, Dortmund University of Technology) – Statistical Learning
- Prof. Scheffer – (Department of Computer Science, University of Potsdam) – Bioinformatics for HIV
- Prof. Shafer (Devision of Infectious Diseases, Stanford University) – Bioinformatics for HIV
- Prof. Shioda (Research Institute of Microbial Diseases, Osaka University, Department of Viral Infections) – Bioinformatics for HIV
- Prof. Sönnerborg (Karolinska Institute) – Bioinformatics for HIV
- Dr. Thomas (Max Planck Institute for Neurological Research, Cologne) – Genomic Aberrations in Cancer
- Prof. Überla (Molecular and Medical Virology, Ruhr-University Bochum) – Bioinformatics for HIV
- Dr. Walter (Institute of Clinical and Molecular Virology, German National Reference Centre for Retroviruses) – Bioinformatics for HIV
- Prof. Weickert (Computer Science, Saarland University) – Protein Structure
- Prof. Wenz (Chemistry, Saarland University) – Cheminformatics
- Dr. Xenarios (Swiss Institute for Bioinformatics, Lausanne) – GISAID
- Prof. Zazzi (Virology, University of Siena) – Bioinformatics for HIV

We are collaborating with the following scientific consortia:

Arevir Consortium – HIV Bioinformatics

BioSapiens - European Network of Excellence for Genome Annotation

CancerDIP Consortium – Epigenetics
CHAIN Consortium - EU Project on HIV Resistance
DFG Clinical Research Group 129 – HCV Bioinformatics
EUResist – EU STREP on HIV resistance analysis
Fraunhofer-Max Planck Cluster on Machine Learning – Computational Biology
GISAID Consortium – Initiative on Sharing Influenza Data
HIV-GRADE Consortium – HIV Bioinformatics for German Patients
Oncogene Consortium - Analysis of Genomic Data Pertaining to Cancer

We are cooperating with the following companies:

BioSolveIT GmbH, Sankt Augustin (`http://www.biosolveit.de`, Dr. Holger Claußen, Dr. Sally Hindle, Dr. Christian Lemmen), a startup company of which Thomas Lengauer is a co-founder – Docking and Drug Screening
Pfizer Pharma GmbH (Germany) – HIV Coreceptor Usage
Virco BVBA (Mechelen, Belgium) and VircoLab, Inc. (Durham, NC, USA) – Bioinformatics for HIV

Software and Web Services

It is the acclaimed intention of the Department to develop bioinformatics software that is useful to a wide user community. Thus the Department also commits to realizing channels of dissemination and evaluation of the software, be they commercial or non-commercial. Most of our tools, such as the analysis software of HIV genotypes for drug resistance (Geno2Pheno Server) or the (epi-) genome browser (EpiGRAPH) have been made accessible non-commercially via an Internet server offer. Some software predating the developments that took place at this Institute has mostly been made available commercially through the startup company BioSolveIT, Sankt Augustin (`www.biosolveit.de`), co-founded by Thomas Lengauer in 2001. This comprises especially software on biomolecular docking and drug screening.

Within the report period, several new software packages as well as web services have been made available via our web server (see `http://www.mpi-inf.mpg.de/departments/d3/software.html`). These include:

BioMyn (`http://www.biomyn.de/`) is a comprehensive online resource that integrates information related to human genes and proteins from over a dozen external databases. It includes Gene Ontology annotations of human genes and proteins, sequence family classifications, protein domain architectures, metabolic and signaling pathways, protein interactions and protein complexes, and disease associations.

DASMI (`http://www.dasmi.de/`) uses the DAS protocol to share experimental and predicted molecular interaction data [3]. The current DASMIweb client (`http://www.dasmiweb.de/`) is focused on interactions of human proteins and of Pfam domains. It also supports scoring the retrieved interactions with confidence values using different quality measures.

DomainGraph (http://domaingraph.bioinf.mpi-inf.mpg.de/) is a Java plugin for Cytoscape, a free open-source software platform for visualization and analysis of biomolecular networks (http://www.cytoscape.org). DomainGraph decomposes proteins into their constituent domains and generates a network of interacting protein domains [9]. In particular, it allows the integration of exon expression data, which supports the analysis of alternative splicing events and the characterization of their effects on protein and domain interaction networks.

EpiGRAPH (http://epigraph.mpi-inf.mpg.de/) enables biologists to analyze genome and epigenome datasets with powerful statistical and machine learning methods. In a typical workflow, the user uploads a set of genomic regions of interest (e.g. experimentally mapped enhancers, hotspots of epigenetic regulation or sites exhibiting disease-specific alterations), and EpiGRAPH searches a large database of (epi-) genomic attributes for significant overlap and correlation with the regions in the input dataset. Furthermore, EpiGRAPH can predict the status of genomic regions that were not included in the input dataset.

FunSimMat (http://www.funsimmat.de/) is a comprehensive functional similarity database that provides several different semantic similarity measures for Gene Ontology terms [15]. It offers various precomputed functional similarity values for proteins contained in UniProt Knowledge Base and for protein families and domains in the Pfam and SMART databases. The web interface enables users to efficiently perform both semantic similarity searches with GO terms and functional similarity searches with proteins or protein families.

Galinter (http://galinter.bioinf.mpi-inf.mpg.de/) is a method for aligning non-covalent interactions between different protein-protein interfaces [20]. The method aligns the vector representations of van der Waals interactions and hydrogen bonds based on their geometry. A score based on a statistical model is provided for assessing the significance of a match between two interfaces.

GOdot (http://godot.bioinf.mpi-inf.mpg.de/) The GOdot method assesses the variability of protein function in protein sequence and structure space [18]. Various regions in this space exhibit considerable difference in the local conservation of molecular function. GOdot analyzes and captures local function conservation by means of logistic curves. These provide a simple yet consistent statistical model for the complex relations between protein sequence, structure, and function. Based on the analysis, the GOdot web-server predicts molecular function of a query protein with known structure but unknown function.

MethMarker (http://methmarker.mpi-inf.mpg.de/) facilitates the design of DNA methylation assays covering different experimental technologies including COBRA, bisulfite SNuPE, bisulfite pyrosequencing, MethyLight and MSP. It also implements a systematic workflow for design, optimization and (computational) validation of DNA methylation biomarkers. This workflow starts from a preselected differentially methylated region (DMR) and results in an optimized DNA methylation assay that is ready to be tested in a large-scale clinical trial.

RTREEMIX is an easy-to-use R package for estimating evolutionary pathways and genetic progression scores[8]. The package is of use in analyzing evolutionary processes in diseases such as cancer and HIV. It is based on efficient C/C++ code provided by a modified version of the MTREEMIX software, which is available as a free R package. The package implements the main functionality of MTREEMIX for model fitting and adds new functions for estimating genetic progression scores with corresponding confidence intervals and for performing stability analysis on different levels of the fitted models. Moreover, the package offers various diagnostic tools and functions for visualization, for example, plotting the estimated mixture models.

Among the new packages, the most requested ones are FUNSIMMAT (with 1.4 million queries since November 2007[1], EPIGRAPH (with hundreds of users even before its publication date) and RTREEMIX (with over a thousand downloads since its release in October 2008). Among the previously reported packages GENO2PHENO continues to be extremely popular with over 100,000 hits for GENO2PHENO[coreceptor] since June 2004 and over 40,000 hits for GENO2PHENO[resistance] since December 2000. Other previously reported packages that are well accepted by the community include BIQ ANALYZER, which became a standard tool for processing DNA methylation data with dozens of analyses carried out each day, NETWORK-ANALYZER for computing topological network parameters with about 2,500 downloads since January 2006, and TOPGO for Gene Ontology term enrichment computation with about 4,900 downloads during the last year.

We will continue to target much of our future software to larger user communities and exploit the distribution channels open to us. This requires special measures for the hardware and software infrastructure for the Department. We have allocated a full scientist (Joachim Büch, see Section 29.10) and hired student support to the responsibility for installing and maintaining an operative environment for local bioinformatics software and databases and the servers offered to the community. This position is responsible for maintaining continuity when developers leave the Department and for keeping the hard- and software up-to-date. The developers support this effort by bringing the software to a mature beta-test stage and help with setting up the routine software configuration.

Prizes and Awards

Thomas Lengauer has been elected member of the German Academy of Science and Engineering - acatech. Christoph Bock has been awarded an Otto-Hahn Medal in the year 2009 for his dissertation. Christoph Bock has received a Feodor Lynen Fellowship by the Alexander von Humboldt Foundation to pursue postdoctoral research in Harvard University and MIT Broad Institute. Dorothea Emig has received a travel grant from the Boehringer Ingelheim Fonds (Foundation for Basic Research in Medicine) for a research visit at the University of California, Berkeley, USA. Christoph Hartmann has received a DAAD travel grant for a research visit at the University of California, San Francisco, USA. Konstantin Halachev, Lars Feuerbach, and Yassen Assenov have received EU travel grants (Marie Curie – Genome Architecture in Relation to Disease) to attend MC-CARD conferences and workshops.

[1] This high number of hits results partly from whole-genome screens that are run by several users. We have had over 190 different users of FunSimMat.

References

- [1] A. Altmann, N. Beerenwinkel, T. Sing, I. Savenkov, M. Däumer, R. Kaiser, S.-Y. Rhee, W. J. Fessel, R. W. Shafer, and T. Lengauer. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antiviral Therapy*, 12(2):169–178, 2007.
- [2] A. Altmann, M. Däumer, N. Beerenwinkel, Y. Peres, E. Schülter, J. Büch, S.-Y. Rhee, A. Sönnerborg, W. J. Fessel, R. W. Shafer, M. Zazzi, R. Kaiser, and T. Lengauer. Predicting the response to combination antiretroviral therapy: Retrospective validation of geno2pheno-THEO on a large clinical database. *The Journal of Infectious Diseases*, 199:999–1006, 2009.
- [3] H. Blankenburg, R. D. Finn, A. Prlić, A. M. Jenkinson, F. Ramírez, D. Emig, S.-E. Schelhorn, J. Büch, T. Lengauer, and M. Albrecht. DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10):1321–1328, 2009.
- [4] C. Bock, K. Halachev, J. Büch, and T. Lengauer. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. *Genome Biology*, 10:R14, 2009.
- [5] C. Bock and T. Lengauer. Computational epigenetics. *Bioinformatics*, 24(1):1–10, 2008.
- [6] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. CpG island mapping by epigenome prediction. *PLoS Computational Biology*, 3(6):1055–1070, 2007.
- [7] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Research*, 36(10):e55, 2008.
- [8] J. Bogojeska, A. Alexa, A. Altmann, T. Lengauer, and J. Rahnenführer. Rtreemix: an R package for estimating evolutionary pathways and genetic progression scores. *Bioinformatics*, 24(20):2391–2392, 2008.
- [9] D. Emig, M. S. Cline, T. Lengauer, and M. Albrecht. Integrating expression data with domain interaction networks. *Bioinformatics*, 24(21):2546–2548, 2008.
- [10] C. Hartmann, I. Antes, and T. Lengauer. Docking and scoring with alternative side-chain conformations. *Proteins: Structure, Function, and Bioinformatics*, 74(3):712–726, 2009.
- [11] J. Kamradt, V. Jung, K. Wahrheit, L. Tolosi, J. Rahnenführer, M. Schilling, R. Walker, S. Davis, M. Stöckle, P. Meltzer, and B. Wullich. Detection of novel amplicons in prostate cancer by comprehensive genomic profiling of prostate cancer cell lines using oligonucleotide-based arrayCGH. *PLoS ONE*, 2(8):e769, 2007.
- [12] T. Lengauer, O. Sander, S. Sierra, A. Thielen, and R. Kaiser. Bioinformatics prediction of HIV coreceptor usage. *Nature Biotechnology*, 25(12):1407–1410, 2007.
- [13] T. Mikeska, C. Bock, O. El-Maarri, A. Hübner, D. Ehrentraut, J. Schramm, J. Felsberg, P. Kahl, R. Büttner, T. Pietsch, and A. Waha. Optimization of quantitative MGMT promoter methylation analysis using pyrosequencing and combined bisulfite restriction analysis. *Journal of Molecular Diagnostics*, 9(3):368–381, 2007.
- [14] M. Rosen-Zvi, A. Altmann, M. Prosperi, E. Aharoni, H. Neuvirth, A. Sönnerborg, E. Schülter, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi, and T. Lengauer. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. In *Proceedings of the 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2008)*, Toronto, Canada, 2008, *Bioinformatics*, vol. 24, pp. i399–406. Oxford University.
- [15] A. Schlicker and M. Albrecht. FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Research*, 36(Database Issue):D434–D439, 2008.

- [16] W. A. Schulz, A. Alexa, V. Jung, C. Hader, M. J. Hoffmann, M. Yamanaka, S. Fritzsche, A. Wlazlinski, M. Müller, T. Lengauer, R. Engers, A. R. Florl, B. Wullich, and J. Rahnenführer. Factor interaction analysis for chromosome 8 and DNA methylation alterations highlights innate immune response suppression and cytoskeletal changes in prostate cancer. *Molecular Cancer*, 6:1–16, 2007.
- [17] T. Sing, A. J. Low, N. Beerenwinkel, O. Sander, P. K. Cheung, F. S. Domingues, J. Büch, M. Däumer, R. Kaiser, T. Lengauer, and P. R. Harrigan. Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antiviral Therapy*, 12(7):1097–1106, 2007.
- [18] N. Weinhold, O. Sander, F. S. Domingues, T. Lengauer, and I. Sommer. Local function conservation in sequence and structure space. *PLoS Computational Biology*, 4:e1000105, 2008.
- [19] C. Welsch, F. S. Domingues, S. Susser, I. Antes, C. Hartmann, G. Mayr, A. Schlicker, C. Sarrazin, M. Albrecht, S. Zeuzem, and T. Lengauer. Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of HCV. *Genome Biology*, 9(1):R16, 2008.
- [20] H. Zhu, I. Sommer, T. Lengauer, and F. S. Domingues. Alignment of non-covalent interactions at protein-protein interfaces. *PLoS ONE*, 3(4):e1926, 2008.

4 The Computer Graphics Group (D4)

History

The computer graphics group was established in 1999 with the appointment of Hans-Peter Seidel. The group currently consists of about 30 researchers. Over the last decade more than 20 former group members [1] received offers for tenured faculty positions, and the group graduated 35 PhD students.

Vision and Research Directions

During the last three decades computer graphics established itself as a core discipline within computer science and information technology. Ten years ago, most digital content was textual. Today it has expanded to include audio, images, video, and a variety of graphical representations. New and emerging technologies such as multimedia, digital television, telecommunication and telepresence, virtual reality, or 3D-internet further indicate the potential of computer graphics in the years to come. Typical for the field is the coincidence of very large data sets with the demand for fast, and possibly interactive, high quality visual feedback. Furthermore, the user should be able to interact with the environment in a natural and intuitive way.

In order to address the challenges mentioned above, a new and more integrated scientific view of computer graphics is required. In contrast to the classical approach to computer graphics which takes as input a scene model – consisting of a set of light sources, a set of objects (specified by their shape and material properties), and a camera – and uses simulation to compute an image, we like to take the more integrated view of *3D Image Analysis and Synthesis* for our research. We consider the whole pipeline from data acquisition, over data processing to rendering in our work. In our opinion, this point of view is necessary in order to exploit the capabilities and perspectives of modern hardware, both on the input (sensors, scanners, digital photography, digital video) and output (graphics hardware, multiple platforms) side. Our vision and long term goal is the development of methods and tools to efficiently handle the huge amount of data during the acquisition process, to extract structure and meaning from the abundance of digital data, and to turn this into graphical representations that facilitate further processing, rendering, and interaction.

[1] T. Ertl (C4, Stuttgart), P. Slusallek (C4, Saarbrücken), G. Greiner (C4, Erlangen), L. Kobbelt (C4, Aachen), W. Heidrich (Full Prof.,Univ. British Columbia, Canada), R. Westermann (C4, Munich), A. Kolb (C4, Siegen), S. Gumhold (W3, Dresden), M. Magnor (W3, Braunschweig), H. Theisel (W3, Magdeburg), B. Rosenhahn (W3, Hannover), M. Botsch (W3, Bielefeld) H. Lensch (W3, Ulm), M. Stamminger (C3, Erlangen), V. Blanz (W2, Siegen), A. Belyaev (Reader, Edinburgh, UK) J. Kautz (Senior Lecturer, UCL, London), P. Bekaert (C3, Univ. Limburg, Belgium), C. Soler (INRIA Rhone Alpes, France), R. Zayer (INRIA Lorraine, France)

In order to make progress along the lines above, our work is both theoretical and practical with a focus on first-class research on new methods and algorithms, as well as on the integration of new algorithms into functioning software systems, and the experimental validation of systems in specific application scenarios that are of practical relevance. We also try to provide a stimulating environment for junior researchers that allows them to develop and build their own research programs and groups.

Research Areas and Structure of the Group

As mentioned above, we consider the whole pipeline from data acquisition over data processing to rendering in our work. Within this framework our choice of research areas is long-term. We reconsider them, as senior researchers leave and as new opportunities arise. Hiring decisions on all levels (PhD students, postdocs, research associates) are made on quality and fit into our research program. Our research is currently organized into the following eight research areas, each having its own small group of coordinators:

- Digital Geometry Processing (M. Wand)
- Visualization (T. Schultz and M. Sips)
- Integrative Scientific Computing (R. Strzodka)
- Markerless Motion Capture and Multiview Stereo Processing (B. Rosenhahn and T. Thormählen)
- Multimedia Information Retrieval (M. Müller)
- General Appearance Acquisition and Computational Photography (H. Lensch)
- Advanced Global Illumination and Realtime Realistic Image Synthesis (E. Eisemann, T. Grosch and K. Myszkowski)
- High Dynamic Range Imaging and Perception Issues in Graphics (K. Myszkowski)

However, we are not organized into disjoint subgroups, and there is little hierarchy. While each of the areas has its specific focus, some of them also have significant overlaps. Likewise, the students and researchers working in each area are dynamically formed teams rather than specifically dedicated staff. The area coordinators together with Hans-Peter Seidel serve as an internal steering committee for the group. They also play the role of advisors or co-advisors of various doctoral students.

Some Achievements

We have been pursuing first-class research along our long-term research agenda, and members of the group have actively published in the top conferences and journals (see Section 30.17 for details). In 2008, e.g., the group published 8 papers at ACM SIGGRAPH (and a total of

11 papers in ACM TOG), 5 papers at EUROGRAPHICS, and 6 papers at IEEE CVPR - all in a single year. In addition, we also won the best paper award at IEEE Visualization.

We have actively participated in program committees and have given numerous invited talks and tutorial presentations at major national and international events (see Section 30.13 for details). Our software has been successfully integrated and validated in a variety of projects (see Sections 30.12 and 30.16), and researchers from the group have spread out to other institutions.

In the following we briefly highlight some of our most exciting results:

Linear angle based parameterization (SGP '07) One of the most prominent approaches for mesh parametrization is Angle Based Flattening (ABF), which directly formulates the problem as a constrained nonlinear optimization in terms of angles. Since the original formulation of ABF, a steady research effort has been dedicated to improving its efficiency. We have developed a novel linear version of angle based flattening, based on a reformulation of the problem using the notion of estimation error. The error induced by this linearization is quadratic in terms of the error in angles. Besides performance speedup, the simplicity of the current setup makes re-implementation and reproduction of our results straightforward. (*Investigator: Rhaleb Zayer)*

Adaptive feature-preserving non-local denoising of static and time-varying range data (CAD '08) We present a novel method for noise removal on static and time-varying range data. Our approach predicts the restored position of a perturbed vertex using similar vertices in its neighborhood. It defines the required similarity measure in a new non-local fashion that compares regions of the surface instead of point pairs. This allows our algorithm to obtain a more accurate denoising result than previous state-of-the-art approaches and, at the same time, to better preserve fine features of the surface. Another interesting component of our method is that the neighborhood size is not constant over the surface but adapted close to the boundaries which improves the denoising performance in those regions of the dataset. (*Investigators: Oliver Schall and Alexander Belyaev)*

Isometric Registration of Ambiguous and Partial Surface Data (CVPR '09) We introduce a new shape matching algorithm for computing correspondences between 3D surfaces that have undergone (approximately) isometric deformations. The new approach makes two main contributions: First, the algorithms is, unlike previous work, robust to topological noise, such as large holes or false connections, which are both observed frequently in real-world scanner data. Second, the algorithm samples the space of feasible solutions such that uncertainty in matching can be detected explicitly. We employ a novel randomized feature matching algorithm in order to find robust subsets of geodesics to verify isometric consistency. (*Investigators: Art Tevs and Michael Wand)*

Estimating Crossing Fibers in Diffusion-Weighted MRI-Imaging: A Tensor Decomposition Approach (IEEE Visualization '08) Diffusion weighted magnetic resonance imaging (DW-MRI) is a unique tool for non-invasive investigation of major nerve fiber tracts. Many types of features in DW-MRI data require the estimation of the number, orientations, and volume

fractions of individual nerve fiber tracts within a voxel. Our contribution improves the reliability of such estimates in cases where high angular resolution data is used to investigate crossing fiber bundles. This is achieved by a low-rank approximation of higher order tensors. These are used to model the orientation distribution functions which stem from Q-Ball imaging and spherical deconvolutions. (*Investigator: Thomas Schultz)*

3D-Modeling by Ortho-Image Generation from Image Sequences (ACM SIGGRAPH '08) We develop a semi-automatic approach for the generation of a high-quality 3D model of a static object from an image sequence that was taken by a moving, uncalibrated consumer camera. First, the camera parameters for each input image are estimated by automatic camera tracking. Afterwards, techniques from image-based rendering are used to generate an orthographic projection on a bounding box that is placed around the object. These ortho-images can be imported into any modeling package. The approach is capable of handling not only diffuse surfaces, but even translucent or specular surfaces. (*Investigator: Thorsten Thormählen)*

Drift-free tracking of rigid and articulated objects (CVPR '08) Relying only on image features that are tracked over time does not prevent the accumulation of small errors which results in a drift away from the target object. The error accumulation becomes even more problematic in the case of multiple moving objects due to occlusions. To solve the drift problem for tracking, we propose an analysis-by-synthesis framework that uses reference images to correct the pose. It comprises occlusion handling and is successfully applied to crash test video analysis. (*Investigators: Jürgen Gall and Bodo Rosenhahn)*

Optimization and Filtering for Human Motion Capture - A Multi-Layer Framework (Int. J. Comp. Vision '09) Typical cues for motion capture are silhouettes, edges, color, motion, and texture. In general, a multi-cue integration is necessary for tracking complex objects. We propose an adaptive weighting scheme that combines complementary cues, namely silhouettes on one side and optical flow as well as local descriptors on the other side. We introduce a multi-layer framework that combines stochastic optimization, particle filtering, and local optimization. While the first layer relies on interactive simulated annealing, the second layer refines the estimates by filtering and local optimization such that the accuracy is increased and ambiguities are resolved over time without imposing restrictions on the dynamics. (*Investigator: Jürgen Gall and Bodo Rosenhahn)*

Performance Capture from Sparse Multi-view Video (ACM SIGGRAPH '08) Since most marker-based and marker-free motion capture systems measure the motion in terms of a kinematic skeleton, they have to be combined with other scanning technologies to capture the time-varying shape and surface details of the human body surface. However, dealing with people wearing arbitrary clothing from only video streams is still not possible. To overcome this limitation, we propose a framework, enabling the direct animation of a high-quality static human scan from unaltered video footage. The proposed method explicitly abandons any traditional skeletal or motion parameterization and poses performance capture as deformation capture. Our algorithms successfully capture motion and time-varying shape detail even

of people wearing wide apparel, while preserving its spatio-temporal coherence over time. (*Investigators: Edilson de Aguiar, Naveed Ahmed, Carsten Stoll)*

Fitting a Morphable Model to 3D Scans of Faces (ICCV '07) We present a top-down approach to 3D data analysis by fitting a morphable model to scans of faces. In a unified framework, the algorithm optimizes shape, texture, pose and illumination simultaneously. The algorithm can be used as a core component in face recognition from scans. In an analysis-by-synthesis approach, raw scans are transformed into a PCA-based representation that is robust with respect to changes in pose and illumination. Illumination conditions are estimated in an explicit simulation that involves specular and diffuse components. (*Investigators: Kristina Scherbaum, Volker Blanz)*

Relighting Objects from Image Collections (CVPR '09) We propose a novel method to simultaneously recover surface geometry, reflectance properties of opaque objects, and prevailing lighting conditions at the time of image capture from just a small number of input photographs. While there exist previous approaches to recover reflectance properties, our system is the first to work on images taken under almost arbitrary, changing lighting conditions. This enables us to use images we took from a community photo collection website. (*Investigators: Christian Fuchs and Hendrik Lensch)*

Fluorescent Immersion Range Scanning (ACM SIGGRAPH '08) High-quality 3D scanning of optically challenging materials calls for novel acquisition techniques. We introduce an extension of a traditional laser scanning setup that allows for robust surface scanning of objects made of translucent, transparent, glossy, and very dark materials. The key idea is to observe the rays of the structured light source before they reach the object, and to disregard any light that is being reflected from the surface. We achieve this by embedding our object in a participating medium, in particular a fluorescent dye in water, which scatters incoming light rays only once. By performing a simple thresholding on the space-time gradient, we obtain surface geometry of many different classes of materials at a high resolution and low noise figures. (*Investigators: Matthias Hullin, Martin Fuchs, Ivo Ihrke, Hendrik Lensch)*

Adaptive Sampling of Reflectance Fields (ACM TOG '07) Image-based relighting achieves high quality in rendering, but it requires a large number of measurements of the reflectance field. We propose novel sampling techniques that improve on the trade-offs between measurement effort and reconstruction quality. Specifically, we (i) demonstrate that sampling with point lights and from a sparse set of incoming light directions creates artifacts which can be reduced significantly by employing extended light sources for sampling, (ii) propose a sampling algorithm which incrementally chooses light directions adapted to the properties of the reflectance field being measured, thus capturing significant features faster than fixed-pattern sampling, and (iii) combine reflectance fields from two different light domain resolutions. (*Investigators: Martin Fuchs, Volker Blanz, Hendrik Lensch)*

Towards Passive 6D Reflectance Field Displays (ACM SIGGRAPH '08) Traditional flat screen displays present 2D images. We explore different designs of multi-dimensional displays

which passively react to the light of the environment behind. The prototypes physically implement a reflectance field and generate different light fields depending on the incident illumination, for example light falling through a window. We discretize the incident light field using an optical system, and modulate it with a 2D pattern, creating a flat display which is view *and* illumination-dependent. (*Investigators: Martin Fuchs, Hendrik Lensch)*

Render2MPEG: A Perception-based Framework Towards Integrating Rendering and Video Compression (EG '08) Currently, 3D animation rendering and video compression are completely independent processes, even if rendered frames are streamed on-the-fly within a client-server platform. We present a novel framework where the renderer and MPEG codec are coupled through a straightforward interface that provides precise motion vectors from the rendering side to the codec and perceptual error thresholds for each pixel in the opposite direction. The availability of the discrete cosine transform coefficients at the codec side enables to use advanced models of the human visual system in the perceptual error threshold derivation without incurring any significant cost. Those error thresholds are used to control the rendering quality and make it well aligned with the compressed stream quality. (*Investigators: Robert Herzog, Karol Myszkowski)*

Imperfect Shadow Maps for Efficient Computation of Indirect Illumination (ACM TOG '08) We present a method for interactive computation of indirect illumination in large and fully dynamic scenes based on approximate visibility queries. While the high-frequency nature of direct lighting requires accurate visibility, indirect illumination mostly consists of smooth gradations, which tend to mask errors due to incorrect visibility. We exploit this by approximating visibility for indirect illumination with low-resolution shadow maps rendered from a crude point-based representation of the scene. We demonstrate that imperfect shadow maps are a valid approximation to visibility, which makes the simulation of global illumination an order of magnitude faster than using accurate visibility. (*Investigators: Tobias Ritschel, Thorsten Grosch)*

Real-Time, All-Frequency Shadows in Dynamic Scenes (ACM SIGGRAPH '08) In order to achieve real-time, photo-realistic rendering of computer-generated scenes, we introduce the convolution soft shadow method. It is a very fast method for rendering plausible soft shadows based on convolution theory. It requires only a constant-time memory lookup, thereby enabling us to render soft shadows at hundreds of frames per second for a single area source. Environment-lit scenes can be rendered from a collection of approximating area light sources. Even though shadows are only approximate, the results are virtually indistinguishable from reference renderings and are produced at real-time frame rates. (*Investigators: Thomas Annen, Zhao Dong)*

Eikonal Rendering: Efficient Light Transport in Refractive Objects (ACM SIGGRAPH '08) We develop a new method for real-time rendering of sophisticated lighting effects in and around refractive objects. This enables us to realistically display refractive objects with complex material properties. User-controlled changes of lighting positions only require a few seconds of update time. Our method is based on a set of ordinary differential equations derived

from the eikonal equation, the main postulate of geometric optics. This set of equations allows for fast casting of bent light rays with the complexity of a particle tracer. Based on this concept, we also propose an efficient light propagation technique using adaptive wavefront tracing. Efficient GPU implementations for our algorithmic concepts enable us to render complex visual effects that were previously not reproducible in real-time. (*Investigators: Ivo Ihrke, Gernot Ziegler, Art Tevs)*

3D Unsharp Masking for Scene Coherent Enhancement (ACM SIGGRAPH '08) We present a new approach for enhancing local scene contrast by unsharp masking over arbitrary surfaces under any form of illumination. Our adaptation of a well-known 2D technique to 3D interactive scenarios is designed to aid viewers in tasks like understanding complex or detailed geometric models, medical visualization, and navigation in virtual environments. Our holistic approach enhances the depiction of various visual cues, including gradients from surface shading, surface reflectance, shadows, and highlights, to ease estimation of viewpoint, lighting conditions, shapes of objects and their world-space organization. Motivated by recent perceptual findings on 3D aspects of the Cornsweet illusion, we create scene coherent enhancements by treating cues in terms of their 3D context. (*Investigators: Tobias Ritschel, Kaleigh Smith, Thorsten Grosch, Karol Myszkowski)*

Dynamic Range Independent Image Quality Assessment (ACM SIGGRAPH '08) The diversity of display technologies and introduction of high dynamic range imagery causes the necessity of comparing images of radically different dynamic ranges. Current quality assessment metrics are not suitable for this task. We present a novel image quality metric capable of operating on an image pair where both images have arbitrary dynamic ranges. Our metric utilizes a model of the human visual system, and its central idea is a new definition of visible distortion based on the detection and classification of visible changes in the image structure. Our metric is carefully calibrated and its performance is validated through perceptual experiments. (*Investigators: Tunc Aydin, Rafal Mantiuk, Karol Myszkowski)*

Prizes and Awards

Several current and former group members have received awards for their work during the reporting period. Hendrik Lensch received an Emmy-Noether-Fellowship from the German Research Foundation, and Bodo Rosenhahn won the DAGM Olympus Prize 2007. Michael Goesele (now with TU Darmstadt) and Christian Theobalt (now with Stanford) both received the Eurographics Young Researcher Award. Ivo Ihrke received the Eduard Martin Prize 2007 (PhD award by Saarland University). Edilson de Aguiar (Disney/CMU), Tongbo Chen (USC), Jürgen Gall (ETH), Ivo Ihrke (Alexander von Humboldt Foundation, UBC), Andrei Lintu (INRIA Lorraine), Rafal Mantiuk (UBC) and Thomas Schultz (DAAD, U. Chicago) all received postdoc fellowships. Finally, Thomas Schultz, Jürgen Gall, Oliver Schall, Grzegorz Krawczyk, Kaleigh Smith and Zhao Dong all received best paper awards at major international conferences.

Projects and Cooperations

The group is part of the Cluster of Excellence on "Multimodal Computing and Interaction" that was established within the framework of the German Excellence Initiative. Hans-Peter Seidel is the scientific coordinator of the cluster.

In addition to the collaborations within the Cluster of Excellence, within the institute (surface reconstruction and shape matching, molecular dynamics on graphics hardware, music IR), and within the university (ray tracing, interactive global illumination, mesh processing, 3D video processing, speech synchronization, synthetic virtual characters), and in addition to numerous collaborations between individual researchers (including several of our former graduates), we have also established a number of formal cooperations with other institutions on both the national and international level. On a European level, these include the EU projects Aim@Shape (NoE), and 3DTV (NoE). We also collaborate directly with Daimler (crash test video analysis), MERL (passive reflectance field displays), BrightSide Technologies/Dolby (high dynamic range video compression), and Samsung (hybrid rendering). Details can be found in Section 30.16.

Max Planck Center for Visual Computing and Communication

The Max Planck Center for Visual Computing and Communication (MPC-VCC) with corresponding research activities at the Max Planck Institute for Computer Science and at Stanford University was established jointly by MPG and Stanford University in October 2003. The proposed collaboration has two intertwined goals: Establish a joint research program in the area of visual computing and communication and incorporate a strong career development component to alleviate the shortage of qualified faculty and scientists in information technology in Germany.

The Max Planck Center fosters the professional development of a small number of selected outstanding individuals by providing them with the opportunity to work at Stanford University as visiting assistant professors in the area of visual computing and communication for two years and then return to Germany to continue their research as leader of a junior research group at the Max Planck Institute for Computer Science and ultimately as a professor at a German university.

The center is directed jointly by Hans-Peter Seidel (MPII) and Bernd Girod (Stanford).

Group Development

Overall, the structure of the group has been essentially stable during the reporting period, but there has again been considerable fluctuation in the individual composition of the group.

Elmar Eisemann (INRIA Rhone-Alpes, 2008), Meinard Müller (Univ. Bonn, 2007), Robert Strzodka (Stanford, 2007), Thorsten Thormählen (Adelaide, Australia 2007) and Michael Wand (Stanford, 2007) joined us as senior researchers. Lionel Baboud (INRIA, 2009), Thorsten Grosch (Univ. Koblenz, 2007), Ruxandra Lasowski (TU Munich, 2008), Sung Kil Lee (Postech, Korea, 2009), Makoto Okabe (Univ. Tokyo, 2008), and Mike Sips (Stanford, 2008) joined us as postdocs.

Alexander Belyaev (Edinburgh, UK, 2007), Bodo Rosenhahn (Univ. Hannover, 2008), Hendrik Lensch (Univ. Ulm, 2009), Rhaleb Zayer (INRIA Lorraine), and Hongbo Fu (Assistant Professor, Hong Kong, 2009) all accepted offers for faculty positions and left the group. Michael Goesele who had been doing a postdoc at Univ. Washington returned to Germany and accepted a W1-Professorship at TU Darmstadt.

Edilson de Aguiar (Disney Labs/CMU), Tongbo Chen (USC), Martin Fuchs (Princeton), Jürgen Gall (ETH), Ivo Ihrke (UBC), Andrei Lintu (INRIA Lorraine), Rafal Mantiuk (UBC), Thomas Schultz (Univ. Chicago) and Christian Theobalt (Stanford) accepted offers for postdoc positions at leading institutions abroad. Christian Fuchs (Univ. Ulm) and Waqar Saleem (Univ. Jena) accepted researcher positions at German universities.

Lukas Ahrenberg, Thomas Annen (PDI/Dreamworks, USA), Robert Bargmann, Grzegorz Krawczyk (European Patent Office, Munich), Torsten Langer, Volker Scholz, Kuangyu Shi, Kaleigh Smith (Google Zürich), Wolfram von Funck, Akiko Yoshida (Sharp Research Labs, Japan) all accepted senior positions in industry.

Our current PhD students are Tunc Aydin, Andreas Baak, Piotr Didyk, Zhao Dong, Miguel Granados, Peter Grosche, Nils Hasler, Thomas Helten, Matthias Hullin, Jens Kerber, Verena Konz, Sergey Kosov, Christian Kurz, Tobias Ritschel, Kristina Scherbaum, Mohammed Shaheen, Carsten Stoll, Martin Sunkel, and Art Tevs.

5 The Databases and Information Systems Group (D5)

5.1 Overview

D5 has been established in October 2003. It is headed by Gerhard Weikum, and currently consists of 15 doctoral students and 12 post-doctoral researchers. The group's research aims to bridge the models, algorithmic methods, and architectural paradigms of the three fields of database systems, information retrieval, and data mining. The work is organized into five areas of technical competence:

- knowledge harvesting (coordinated by Gerhard Weikum),
- Web and text mining (coordinated by Srikanta Bedathur),
- ranking and uncertain data management (coordinated by Martin Theobald),
- query processing and optimization (coordinated by Thomas Neumann), and
- distributed data and communities (coordinated by Mauro Sozio).

While each of these areas has its specific focus, they do have thematic overlaps and mutual synergies. Likewise, the students and researchers working in each area are dynamically formed teams rather than specifically dedicated staff. The area coordinators together serve as an internal steering board and individually as communication links between areas. They also play the role of co-advisors or de-facto advisors of various doctoral students.

In addition to the above five areas, D5 intensively cooperates with one of the independent research groups in the Excellence Cluster with focus on:

- efficient search in semistructured data spaces (headed by Ralf Schenkel, see Section 31.9).

5.2 Vision and Research Directions

The group's long-term objective is to develop methodology for knowledge discovery: collecting, organizing, searching, exploring, and ranking facts from structured, semistructured, and textual information sources. Our approach towards this ambitious goal combines concepts, models, and algorithms from several fields, including database systems, information retrieval, statistical learning, and data mining. In the following, we elaborate on this long-term vision, and how it unfolds into technical challenges that drive our research activities.

5.2.1 Long-Term Vision

Today, scientific results are available on the Internet in the form of publications, a main information source for scholars, and in encyclopedias like Wikipedia, a major source for students. Digital libraries and thematic portals combine multiple literature or data collections, but there is neither deep integration nor comprehensive coverage. Information search is limited to keywords and simple metadata, and media like video, images, or speech are most effectively searched by means of manually created annotations.

We envision a comprehensive, multimodal knowledge base of encyclopedic scope but in a formal (machine-processable) representation. This should encompass all human knowledge in terms of explicit facts referring to the concepts and entities of the underlying domains of discourse (e.g., concepts such as enzymes, quasars, or poets and specific entities such as Steapsin, 3C 273, or Bertolt Brecht), to definitions, theorems, and hypotheses, to measurements of natural phenomena and artefacts, and a wealth of video and speech material as well as sensor readings, all of which should be first-class citizens for effective search. In particular, we aim for efficient methods for large-scale information enrichment and knowledge extraction, powerful querying with semantic search capabilities, deployment in a decentralized (peer-to-peer) and social-network environment, and maintenance of information history and knowledge evolution over long time horizons.

A comprehensive knowledge base should know all individual entities of this world (e.g., Nicolas Sarkozy), their semantic classes (e.g., Sarkozy is a Politician), relationships between entities (e.g., Sarkozy presidentOf France), as well as validity times and confidence values for the correctness of such facts. Moreover, it should come with logical reasoning capabilities and rich support for querying and ranking. The benefits from solving this grand challenge would be enormous. Potential applications include but would not be limited to:

- a machine-readable, formalized encyclopedia that can be queried with high precision like a semantic database;
- an enabler for semantic search on the Web, for detecting entities and relations in Web pages and reasoning about them in expressive (probabilistic) logics;
- a backbone for natural-language question answering that would aid in dealing with entities and their relationships in answering who/where/when/etc. questions;
- a key asset for machine translation (e.g., English to German) and interpretation of spoken dialogs, where world knowledge is essential context for disambiguation;
- a catalyst for acquisition of further knowledge and largely automated maintenance and growth of the knowledge base.

With a knowledge base that sublimates the valuable content from the Web, we could address difficult questions that are beyond the capabilities of today's keyword-based search engines. For example, one could ask for drugs that inhibit proteases and would obtain a precise and fairly comprehensive list of drugs for this HIV-relevant family of enzymes. Such advanced information requests are posed by knowledge workers like scientists, students, journalists, historians, or market researchers. Although it is possible today to find relevant answers, this

process is extremely laborious and time-consuming as it often requires rephrasing queries and browsing through many potentially promising but eventually useless result pages. The following example questions illustrate this point:

- *Which German Nobel laureate survived both world wars and outlived all of his four children?* The answer is Max Planck. The bits and pieces for the answer are not that difficult to locate: lists of Nobel prize winners, birth and death dates of these people, family members extracted from biographies, dates of children. Gathering and connecting these facts is straightforward for a human, but it may take days of manually inspecting Web pages.

- *Which politicians are also accomplished scientists?* Today's search engines fail on questions of this kind because they are matching words and returning pages rather than identifying entities like persons and testing their relations. Additionally, the question entails a difficult ranking problem. Wikipedia alone contains hundreds of persons that are listed in the categories Politicians as well as Scientists. An insightful answer must rank important people first, for example, the German chancellor Angela Merkel who has a doctoral degree in physical chemistry, or Benjamin Franklin, and the like.

- *How are Max Planck, Angela Merkel, Jim Gray, and the Dalai Lama related?* The answer is that all four of them have a doctoral degree from a German university, honorary doctorates in the cases of Jim Gray and the Dalai Lama. Discovering interesting facts about multiple entities and their connections on the Web is virtually impossible, by the sheer amount of interconnected pages about the four people.

Note that even though the queries are phrased as natural-language questions, they would remain equally hard if they were expressed in a formal language. Conversely, a rich knowledge base of entities and relations would enable much more effective natural-language question answering.

5.2.2 Research Topics

The outlined long-term vision entails a variety of challenging research directions.

Information enrichment and knowledge harvesting. Starting with simple content representations of text and speech, this work aims to automatically generate semantic representations for concepts, entities, and relations. Methodologically, this should be based on a combination of pattern matching, natural-language processing techniques, and statistical learning. As a second step, relations could be automatically extracted, for example, memberships in organizations, awards of persons, scholarly genealogy, time and place of historical events, protein interactions, etc. A third stage aims to represent the semantics of natural-language sentences (from text or speech sources) in a formal predicate-argument structure such as frames with semantic role labels. Finally, we want to address also the gathering and automatic organization of multimodal information about entities (people, landmarks, etc.) and relations (cultural or sports events, etc.), in the form of photos, videos, or music recordings. At all stages, ontologies can be leveraged as background knowledge;

mapping concepts, entities, and relations to ontology entries and disambiguating them with conflict resolution in the given context is another challenge we plan to address.

Semantic search. The search paradigm on the Web is keyword queries with simple metadata filtering; for structured databases there are richer languages like SQL or XQuery, and the Semantic-Web research advocates languages like SPARQL and OWL. However, none of these expressive languages is able to quantify degrees of (un-)certainty, relevance, or other quality measures in order to rank search results. Future search should be able to reason on information uncertainty in a principled manner, considering different levels of relevance, authority, resolution, completeness, freshness, provenance, etc.

Social networks and human computing. Search capabilities can be further boosted by collaborative data annotation and recommendation via social networks. Such networks are becoming more and more popular, yet it is still widely open how to utilize community behavior towards consistently better search results.

Information history and knowledge evolution. Today, Web repositories capture only current information, but their history would be a rich source of information and latent knowledge. Web archiving bodies like the Internet Archive limit themselves to the preservation of raw data with very little support for querying. The envisioned knowledge base should also be searchable with "time-travel" queries and trend analyses.

Scalable and efficient methods. Often, the huge amounts of data (in the many-Terabytes range) and the computational complexity of certain analyses (e.g., spectral analysis of huge graphs) pose major challenges regarding scalability and efficiency. Distributed platforms like peer-to-peer networks or cloud-computing farms offer great potential for scalability. However, for many of the data-analysis and data-enrichment tasks, decentralized algorithms are largely under-explored or even unknown.

Most of these topics are investigated in our previous and ongoing work; some issues are the subject of planned activities. Note that these topics are orthogonal to the research areas that constitute the structure of D5: the research areas reflect competences, the topics here reflect ongoing and planned activities.

5.3 Achievements

The group is one of the world-wide leading research groups on the integration of database-system (DB) and information-retrieval (IR) methodologies, a direction casually referred to as DB&IR integration. Our most salient, frequently cited contributions are the work on efficient top-k query processing, the work on XML information retrieval that includes the TopX open-source system, and the recent work on knowledge harvesting, search, and ranking centered around the YAGO-NAGA project. Also, the very recent work on scalable indexing and querying of graph-structured RDF data has received very high attention. Our software

systems TopX (on XML IR) and RDF-3X (on RDF search) are among the very best systems of their kinds. In benchmark comparisons, they outperformed all competitors.

It is remarkable that the group has a very successful track record in both communities, DB and IR, with a good number of first-rate publications and impact by citations, open-source software, and specific data collections prepared by the group's research tools. As a syntactic indicator, in the two-year timeframe from spring 2007 through 2009, the group has had 7 full papers in the premier conferences on IR (ACM SIGIR 2007, 3 x ACM SIGIR 2008, ACM CIKM 2008, WWW 2007, WWW 2009) and 9 full papers in the flagship conferences of the DB community (ACM SIGMOD 2008, 2 x ACM SIGMOD 2009, VLDB 2008, ACM PODS 2008, IEEE ICDE 2008, 2 x IEEE ICDE 2009, ACM KDD 2008). Also, the director of D5 gave invited keynotes at SIGMOD 2007 and WSDM 2009, two of the most important conferences in the DB and IR communities. Some of the relatively recent publications have already achieved remarkable citation rates (according to Google Scholar): the WWW 2007 paper on YAGO is cited more than 100 times already; the four papers about our work on efficient top-k query processing (VLDB 2004, SIGIR 2005, VLDB 2005, VLDB 2006) together have around 300 citations.

The group puts major efforts into prototyping software systems, and pursues the philosophy of open-source software dissemination. Currently, the XML IR system TopX, the peer-to-peer search platform Minerva, and the RDF search engine RDF-3X can be freely downloaded from the group's Web site. Other software tools are available on request. Our systems are being used by other projects world-wide. Also, we have a good practice of demonstrating our software in the demo programs of premier conferences (SIGMOD, VLDB, SIGIR, CIDR). These demo programs have become very popular, are fully refereed, and are meanwhile very competitive (sometimes with more than a hundred submissions and acceptance ratios of 25 percent). We have been very successful with demos at the above conferences and also at the Cebit 2009 trade show. Our software systems are also intensively used for new projects within the group itself; for example, TopX is being used for new work on proximity search, Minerva is used for distributed search on audio-visual data, and RDF-3X forms the basis of efficiently querying graph-structured knowledge bases.

With regard to the five research areas that provide the organizational structure of D5, the most salient contributions are the following:

- *Knowledge harvesting (coordinated by Gerhard Weikum)*: using Wikipedia and WordNet as a basis for building one of the world's largest high-quality knowledge bases [15, 16] (see Subsection 31.4.1); high-precision methods for extracting facts from semistructured and textual Web sources, with a novel way of reconciling pattern-based extraction and consistency checking [17] (see Subsection 31.4.2).

- *Web and text mining (coordinated by Srikanta Bedathur)*: an innovative approach to efficiently indexing and querying large archives of Web-history and other collections of versioned or timestamped text documents, with support for "time-travel" keyword search and time-aware ranking [2, 3] (see Subsection 31.5.1); a cutting-edge method for text classification based on logistic regression and variable-length n-grams [4] (see Subsection 31.5.2).

- *Ranking and uncertain data management (coordinated by Martin Theobald)*: leading

algorithms for efficient top-k query processing over semistructured data, and a high-performance engine – the TopX system – for efficient ranked retrieval on large XML data collections [18, 19, 20] (see Subsection 31.6.1); expressive and effective methods for querying graph-structured knowledge bases and ranking the search results, implemented in the NAGA prototype [5, 6, 7] (see Subsection 31.6.2).

- *Query processing and optimization (coordinated by Thomas Neumann)*: leading methods for join-order optimization [8, 9] (see Subsection 31.7.2); the design and development of a scalable engine – coined RDF-3X – for RDF data management that includes novel indexing and query processing methods and leverages the results on join-order optimization [10, 11] (see Subsection 31.7.1).
- *Distributed data and communities (coordinated by Mauro Sozio)*: very general – peer-to-peer-style – distributed algorithms for computing Eigenvector-based authority measures on Web graphs or social networks, with emphasis on autonomy of peers and resilience to dishonest peers [12, 13] (see Subsections 31.8.2 and 31.8.3); a near-optimal approximation algorithm for dynamic replication of data items in an unstructured peer-to-peer network with randomized query routing [14] (see Subsection 31.8.1); cost-beneficial methods for query routing in peer-to-peer networks for keyword search and for publish-subscribe services over distributed text documents, implemented in the Minerva prototype platform [1, 21] (see Subsections 31.8.4 and 31.8.5).

The following paragraphs outline key contributions along our strategic research avenues. We emphasize the open-source software systems that integrate many of our scientific results, serve as experimental platforms and assets for new projects, and are vital components of achieving external impact.

XML Ranked Retrieval and the TopX System. Non-schematic XML data that comes from many different sources and inevitably exhibits heterogeneous structures and annotations (XML tags) cannot be adequately searched using database query languages like XPath or XQuery. Rather the ranked-retrieval paradigm is called for, with relaxable search conditions, various forms of similarity predicates on tags and contents, and quantitative relevance scoring. The group has made highly influential, sustained contributions on ranked retrieval and efficient top-k query processing for textual and semistructured data. Many of the conceptual and algorithmic results have been integrated into the TopX engine for indexing, querying, and ranking large XML collections. TopX has been made available as open-source software. It has been adopted as the reference engine in the popular INEX benchmarking series on XML IR, an annual workshop with more than 50 competing research teams. In the various tasks of the competition itself, TopX has achieved a number of first places over several years. Most notably, TopX has won the INEX 2008 Efficiency Track, outperforming all competitors by a large margin.

Knowledge Harvesting and Discovery by the YAGO-NAGA Project. The recent success of knowledge-sharing communities and the advances in automated information extraction from textual and semistructured Web sources (e.g., Wikipedia) have enabled large-scale harvesting of entity-relationship oriented facts to build a new generation of knowledge

bases. We have developed path-breaking methods to harvest facts from Wikipedia and integrate them with a taxonomic backbone provided by the WordNet thesaurus. The resulting knowledge base, coined YAGO, contains 2 million entities and 20 million facts about them, has an accuracy that is near that of hand-crafted ontologies, and provides a consistent RDF representation of the knowledge. YAGO, including the underlying software tools, is publicly available and has been used in other knowledge-sharing projects including DBpedia. For further growing YAGO from natural-language text sources while retaining its high quality, pattern-based extraction has been combined with logic-based consistency checking in a novel, unified framework coined SOFIE. We have also developed new concepts and algorithms for querying the knowledge base built by the YAGO and SOFIE methodologies, with emphasis on principled ranking using statistical language models. These are implemented in the NAGA search engine. The YAGO-NAGA work has received great attention, is highly cited, and featured in prominent venues like Communications of the ACM.

Scalable Management of RDF Graph Data by the RDF-3X System. In the last few years, there has been rekindled and steadily increasing interest in the Semantic-Web data model RDF that provides flexible representation of subject-property-object triples in a graph-like structure. Important application areas include repositories for computational biology, knowledge-sharing communities, and the interoperability of social networks. Large join queries are an inherent characteristic of searching RDF data, posing a performance challenge already for medium-sized datasets. We have very successfully addressed the issue of scalability, with efficient processing of complex queries on very large databases with billions of RDF triples. This led to the design and implementation of a full-fledged RDF engine coined RDF-3X. Unique features of this system are its streamlined architecture that eliminates the need for physical-design tuning and is well suited for modern hardware, an innovative and aggressive approach to indexing, new algorithms for query processing of complex joins, and a cutting-edge query optimizer that uses a specific form of dynamic programming to determine the best join order for the query execution plan. RDF-3X is available as open-source software and has received great attention. In all published performance studies, it outperforms all other systems by one to two orders of magnitude. We are in the process of using RDF-3X as a basis for several of our internal projects, including the work on comprehensive knowledge bases.

Efficient Indexing, Temporal Querying, and Analysis of Text Archives. Archives of evolving digital content such as blogs, news, or Web pages are becoming available as potential assets for journalists, market and media analysts, intellectual property experts, and many more. To exploit this potential, new forms of access need to be supported, for example, "time-travel" keyword queries, time-aware ranking, identifying the most interesting phrases of particular time period – all in a scalable manner with interactive response times. Our group has been among the very first who addressed this complex of technical problems. The most notable result so far is the TTIX indexing method for time-enriched inverted lists with efficient support for temporal queries. TTIX includes an advanced optimization, based on dynamic programming, for partitioning the overall index by time ranges, in order to minimize the expected processing cost of queries subject to a constraint on the total index size. We

expect that the broader area of scalable text analytics for content archives will soon gain a strategic role in modern data mining and business intelligence.

5.4 Collaborations and Networking

D5 participates in the DFG-funded Excellence Cluster on Multimodal Computing and Interaction (see Section I). Gerhard Weikum is one of the principal investigators and responsible for the work package on comprehensive knowledge bases. The Excellence Cluster is a framework for extensive cooperation on the Saarland University campus: partners are in computational linguistics (Uszkoreit at DFKI, Pinkal and Koller at Saarland University), bioinformatics (Lenhof at Saarland University, Lengauer in the institute), computational logic (Weidenbach in the institute), database systems (Dittrich), data mining (Zeller and Hein at Saarland University, Scheffer in the institute), information retrieval (Müller in the institute), and distributed systems (Maffei at Saarland University, Backes at Saarland University and Max Planck Institute for Software Systems). At this point, these are mostly loose collaborations; joint results (in a subset of these cooperations) are expected in the coming two years.

Within the institute, D5 has primarily been cooperating with D1 (especially Hannah Bast's group) on efficient algorithms for information retrieval. The two groups have jointly participated in the EU-funded Integrated Project DELIS on Dynamically Evolving Large-Scale Information Systems from 2004 until spring 2008. Moreover, they have co-authored several first-rate papers, including a VLDB 2006 paper, a SIGIR 2007 paper, an ICDM 2008 demo paper, and an article in the VLDB Journal in 2008.

Nationally, the group participates in a joint project on scalable middleware for biological network analysis, with two partners at the University of Tübingen (Kaufmann, Kohlbacher) and one at Saarland University (Lenhof). This is in the context of the DFG priority program on Scalable Visual Analytics. Our project within this framework has started in February 2009.

Internationally, the group has been successfully participating in the EU projects AEOLUS, DELIS, DELOS, and SAPIR (see Section 31.13); DELIS and DELOS finished in spring 2008, AEOLUS and SAPIR will be completed in summer 2009. These involved intensive collaborations with the University of Patras, The European Library (TEL) in The Hague, and IBM Research in Haifa, all of which led to joint publications. Two more recently begun EU projects are LiWA (Living Web Archives) and LivingKnowledge, starting in early 2008 and early 2009, respectively. These involve collaboration with major stakeholders on Internet contents, most notably, the European Archive in Amsterdam/Paris (the European sibling of the Internet Archive) and Yahoo! Research in Barcelona.

In addition the group has collaborated, by its own resources without external funding, with the University of Patras (Triantafillou), the University of Athens (Koubarakis), the University of Hong Kong (Mamoulis), the University of Sydney (Fekete), Microsoft Research in Redmond (Chaudhuri, Lomet), Microsoft Research in Cambridge (Vojnovic), the Google Lab in Zurich (Hofmann, Bakir), and Yahoo! Research in Barcelona (Castillo, Donato). All these collaborations produced joint publications, including papers in highly visible and prestigious venues (ACM TOCS, VLDB Journal, KDD, CIKM, CIDR).

The above mentioned collaborations are staggered over several years. At any given point,

about a handful of collaborations is active. The motivation for these projects is twofold. First, personal relations (partly long-term) enable joint work on specific problems that require intensive and focused efforts. The joint work with Microsoft Research is a prototypical example. Second, participation in larger projects is important for networking and can enable access to specific resources or complementary expertise. The joint projects with the European Archive are prototypical examples. The networking also fosters temporary exchange of staff. The group has hosted a fair number of short-term and long-term visitors. Conversely, several D5 members, mostly graduate students, have visited industrial labs for extended periods (Microsoft Research, Google, Telenor).

5.5 Group Development

After the initial build-up of the group until 2005, the group was in steady state from 2005 to 2008, with a healthy balance of graduate students and more experienced researchers. In the last year, the group has undergone noticeable changes, because of three factors: staff fluctuation, a shift in topics, and a growing group size (from 21 in 2007 to 27 in April 2009). The balance between graduate students (15) and postdoctoral researchers (12) is still good, with a recent peak of five newly starting graduate students within a few months.

As for the topic shift, the theme of peer-to-peer search (which was strategic in 2007) is phasing out. On the other hand, the theme of knowledge harvesting is strongly increasing in importance and impact. The same holds, albeit to a lesser extent, for the issues of time-aware search and analyses of digital-content archives. Last but not least, the recent success on scalable management and querying of RDF data has created a substantial momentum in this area, too.

Major *strengths* of D5 are the fact that the group has competences and a very successful track record in both of the two fields of database systems and information retrieval, and the group's combination of system know-how and expertise on algorithmic and mathematical underpinnings. Major *opportunities* on the near-term horizon are applications of the group's open-source software TopX, Minerva, and RDF-3X, and the collaboration with Internet-content organizations like the European Archive. The work on knowledge harvesting and semantic search (originating from the YAGO-NAGA work) is developing strong momentum, both in terms of spawning new projects within the group and being influential externally (by being used by other groups world-wide).

Potential *threats* are the group's diversity of research topics (within a joint overriding theme, however) and the fluctuation of researchers. There are continuous efforts (weekly meetings, reading groups, retreat planned for fall 2009, etc.) to keep the group coherent. Also, the director and the area coordinators pay high attention to keeping the work focused while also being open for new directions that fit into the overall mission. As for staff fluctuation, some of the key members of the group have left during the last two years (Tryfonopoulos who coordinated the peer-to-peer systems work, Schenkel who became the head of an independent research group, Theobald who is now back in a different role, Michel, Ifrim, Luxenburger, Suchanek). Many of these were graduate students, but they played key roles for driving specific sub-projects. On the other hand, although these disruptions are a concern for the group, moving to different career stages and places is certainly healthy for the individuals themselves.

5.6 Prizes and Awards

Gerhard Weikum has become an elected member of the German Academy of Science and Engineering (acatech). Two group members won awards for their outstanding dissertations: Martin Theobald won the Otto-Hahn Medal of the Max-Planck Society and the GI-DBIS Dissertation Award for his work on top-k query processing on text and XML data; he also obtained an Honorable Mention for the ACM SIGMOD Dissertation Award. Sebastian Michel won the Otto-Hahn Medal and the GI-DBIS Dissertation Award for his work on distributed top-k query processing and peer-to-peer search. Two papers of the group won best-paper awards at conferences; one VLDB conference paper was selected for the best-of-conference special issue of the VLDB Journal.

References

- [1] M. Bender, T. Crecelius, S. Michel, and J. X. Parreira. P2P Web search: Make it light, make it fly (demo). In *3rd Biennial Conference on Innovative Data System Research (CIDR 2007)*, Asilomar, USA, 2007, pp. 164–168. www.crdrdb.org.
- [2] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. FluxCapacitor: Efficient time-travel text search. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *33rd International Conference on Very Large Databases (VLDB 2007)*, Vienna, Austria, 2007, pp. 1414–1417. ACM.
- [3] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, eds., *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, Amsterdam, Netherlands, 2007, pp. 519–526. ACM.
- [4] G. Ifrim and G. Weikum. Fast logistic regression for text categorization with variable-length n-grams. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2008)*, Las Vegas, Nevada, USA, 2008, pp. 354–362. ACM.
- [5] G. Kasneci, M. Ramanath, M. Sozio, F. Suchanek, and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *Proceedings of the 25th International Conference on Data Engineering (ICDE 2009)*, Shanghai, China, 2009. IEEE Computer Society.
- [6] G. Kasneci, F. Suchanek, G. Ifrim, S. Elbassuoni, M. Ramanath, and G. Weikum. NAGA: Harvesting, searching and ranking knowledge (demo). In J. Tsong-Li Wang, ed., *Proceedings of the ACM SIGMOD 2008 International Conference on Management of Data (SIGMOD 2008)*, Vancouver, Canada 2008, 2008, pp. 1285–1288. ACM.
- [7] G. Kasneci, F. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. In *24th International Conference on Data Engineering (ICDE 2008)*, Cancun, Mexico, 2008, pp. 953–962. IEEE Computer Society.
- [8] G. Moerkotte and T. Neumann. Dynamic programming strikes back. In J. T.-L. Wang, ed., *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2008)*, Vancouver, Canada, 2008, pp. 539–552. ACM.
- [9] T. Neumann. Query simplification: Graceful degradation for join-order optimization. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD 2009)*, Providence, USA, 2009. ACM.
- [10] T. Neumann and G. Weikum. RDF-3X: a RISC-style engine for RDF. *Proceedings of the VLDB Endowment*, 1(1):647–659, 2008.

- [11] T. Neumann and G. Weikum. Scalable join processing on very large RDF graphs. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD 2009)*, Providence, USA, 2009. ACM.
- [12] J. X. Parreira, C. Castillo, D. Donato, S. Michel, and G. Weikum. The Juxtaposed approximate PageRank method for robust PageRank approximation in a peer-to-peer Web search network. *VLDB Journal*, 17(2):291–313, 2008.
- [13] M. Sozio, T. Crecelius, J. Xavier Parreira, and G. Weikum. Good guys vs. bad guys: Countering cheating in peer-to-peer authority computations over social networks. In *11th International Workshop on the Web and Databases (WebDB 2008)*, Vancouver, Canada, 2008, pp. 103–108. ACM.
- [14] M. Sozio, T. Neumann, and G. Weikum. Near-optimal dynamic replication in unstructured peer-to-peer networks. In M. Lenzerini and D. Lembo, eds., *Proceedings of the Conference on Principles of Database System (PODS 2008)*, Vancouver, Canada, 2008, pp. 281–290. ACM.
- [15] F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 697–706. ACM.
- [16] F. Suchanek, G. Kasneci, and G. Weikum. YAGO - a large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 6(3):203–217, 2008.
- [17] F. Suchanek, M. Sozio, and G. Weikum. SOFIE: A self-organizing framework for information extraction. In *Proceedings of the 18th World Wide Web Conference (WWW 2009)*, Madrid, Spain, 2009. ACM.
- [18] M. Theobald, M. AbuJarour, and R. Schenkel. TopX 2.0 at the INEX 2008 Efficiency Track. In S. Geva, J. Kamps, and A. Trotman, eds., *Preproceedings of the 7th Int. Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, Schloss Dagstuhl, Germany, 2008, pp. 230–244.
- [19] M. Theobald, H. Bast, D. Majumdar, R. Schenkel, and G. Weikum. TopX: Efficient and versatile top-k query processing for semistructured data. *The VLDB Journal*, 17(2):81–115, 2008.
- [20] M. Theobald, R. Schenkel, and G. Weikum. The TopX DB&IR engine (demo). In N. Koudas, ed., *2007 ACM SIGMOD International Conference on Management of Data*, Beijing, 2007, pp. 1141–1143. ACM.
- [21] C. Zimmer, C. Tryfonopoulos, and G. Weikum. Exploiting correlated keywords to improve approximate information filtering. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, eds., *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, 2008, pp. 323–330. ACM.

6 The Automation of Logic Group (RG1)

History

The Automation of Logic Group has been established in September 2005 and is headed by Christoph Weidenbach. The group is following the tradition of the former Programming Logics group with an even stronger focus on the application of the developed results.

Group Composition and Development

The senior scientists and subgroup coordinators are Jörn Freiheit (from the former programming logics department), Viorica Sofronie-Stokkermans, Duc Khanh-Tran, and Uwe Waldmann. Viorica Sofronie-Stokkermans is heading a subgroup on reasoning in complex theories and Uwe Waldmann on first-order model checking. Jörn Freiheit moved to an industry position and Duc Khanh-Tran moved to a tenured position in Vietnam.

There are currently 4 researchers and 9 PhD students in the group. Section 32.1 provides further details to current and recent group members.

Vision and Research Strategy

The vision of the group is to increase the productivity of formal analysis/verification technology through a higher degree of automation of the underlying logics. The following challenges motivate main parts of our work.

- Understand complete and sound reasoning for the combination of theories.
- Drive first-order and propositional reasoning to perfection.
- Understand the mechanics of reasoning with respect to concrete models.
- Show applicability of our methods to reasoning challenges from other areas.
- Scale the applicability of our methods to the size of real world problems.

About two-thirds of our work is of theoretical nature and one third is experimental, in particular on the basis of developed tools. Implementation of our methods is an important part to check their automation potential in practice, to increase impact of our theoretical results by providing our software to other people and eventually to detect further challenges in theory development. For example, our theorem provers SPASS and Waldmeister are among the mostly used systems worldwide in their respective areas demonstrating our impact. Waldmeister even has become a part of Mathematica.

Research Areas and Achievements

Our research is structured under the headings

- First-Order Theorem Proving
- Superposition-Based Inductive Theorem Proving
- Modular Deduction and Verification
- Decision Procedures
- First-Order Model Checking
- Applications
- Software

These headings do not impose a structure on the group. In fact, most of us contribute to several areas and all of us to the development of software. Decision procedures are a typical result of our research in first-order and inductive theorem proving, modular deduction, and model checking.

One syntactic criterion on the success of our work can be seen from CADE-22, the 2009 conference on automated deduction. Among the 32 accepted papers and system descriptions 5 are from our group, meaning that we are the strongest single group at CADE-22. If we also count the contribution of our long term guest Geoff Sutcliffe to CADE-22 there are even 6 contributions from our group.

We will next highlight some of our results of the past two years.

Automated reasoning in combinations of theories (FroCoS 2007, TACAS 2008, KI 2008, AiML 2008, CADE 2009)

Investigators: Carsten Ihlemann, Swen Jacobs, and Viorica Sofronie-Stokkermans

An important part of our research was dedicated to finding methods for efficient reasoning in extensions and combinations of logical theories. We identified and studied the class of *local theory extensions*, for which efficient methods for automated reasoning exist. Our results allow to identify decidable fragments for numerous theories and theory combinations important in mathematics, verification, databases and cryptography, and to provide parameterized complexity results for such theories. See Section 32.5.

Inductive Theorem Proving (CSL 2008, CADE 2009)

Investigators: Matthias Horbach and Christoph Weidenbach

Starting from a superposition calculus for fixed domains, we investigated automated inductive theorem proving. Meanwhile we obtained a number of new decidability results for models represented by atom theories and their generalizations. In addition, we now have proved the first decidability result that relates first-order decidability via saturation with inductive decidability. See Section 32.4.

Combining Large State Spaces with Continuous Flows in Model Checking (ATVA 2007)

Investigators: Swen Jacobs, Willem Hagemann, and Uwe Waldmann

In hybrid systems there is typically a mixture of continuous dynamics and discrete switching. We are investigating new approaches to actually combine the methods for the respective reasoning, in particular with respect to scalability to industrial size problems. See Section 32.7.

Theorem Proving Support for Complex Theories (CAV'09, CADE'09)

Investigators: Carsten Ihlemann, Swen Jacobs, and Viorica Sofronie-Stokkermans

We have an efficient implementation of the method for reasoning in local theory extensions (H-PiLoT) as well as a prototype implementation of an optimized, incremental version (iLoRe). We analyzed the applicability of the methods we developed on various examples from the verification of reactive, real-time and hybrid systems and in mathematics, and (in ongoing work) for efficient reasoning in description logics and the verification of cryptographic protocols. See Section 32.9, Section 32.8.

Feature Modelling (PROSTEP 2008)

Investigators: Christoph Weidenbach and Patrick Wischnewski

A variant of Spass is able to analyze properties of feature models. It scales better than special purpose tools build on top of SAT-solvers, BDD-solvers or constraint approaches and is integrated in the tool VEIA of the company PROSTEP AG. See Section 32.8.

Projects and Cooperations

Together with D1 and colleagues from Saarland University, the group participates in the transregional collaborative research center AVACS. With D1 we investigate applicability of LP technology to the hierarchic combination of linear arithmetic and first-order logic. Together with group of Holger Hermanns from Saarland University we have developed a decomposition approach for enriched probabilistic timed automata. Outside Saarbrücken we have collaborations with the groups of Werner Damm and Ernst-Rüdiger Olderog (University of Oldenburg), and the group of Christoph Scholl (University of Freiburg).

With D5 we are looking at the automation of reasoning for YAGO, see Section 31.4.1, using our first-order theorem proving methods.

With Andrey Rybalchenko from MPI-SWS, our sister institute for software systems, we investigate the integration of our methods into the abstraction refinement approach used in model checking.

For our industrial partner PROSTEP AG we developed a variant of SPASS for feature modelling.

7 The Machine Learning Group (RG2)

History

The Machine Learning Research Group was established in January 2007; it was headed by Tobias Scheffer. The group has been a part of the institute until September 2008; as of October 2008, Tobias Scheffer has accepted a faculty position at the University of Potsdam, Germany. The group's research aimed at understanding how algorithms can be constructed that effectively learn from data to perform better at their tasks. An application-driven line of the group's research was dedicated to applications of machine learning in information retrieval, information security, and healthcare applications.

Group Development

Steffen Bickel, Ulf Brefeld, Michael Brückner, Laura Dietz, Uwe Dick, and Peter Haider joined the Max Planck Institute for Informatics together with Tobias Scheffer in January 2007. Thoralf Klein, Barbara Pogorzelska, Christoph Sawade, Peter Siemen, and Arvid Terzibaschian joined the group as Research Assistants. Ulf Brefeld completed his doctoral degree and accepted a postdoctoral position at Technische Universität Berlin. When Tobias Scheffer joined Potsdam University in October 2008, Laura Dietz decided to join Department 5 while the remaining doctoral students left the institute for Potsdam University.

Vision and Research Strategy

The group's application-oriented line of work was motivated by specific unsolved application problems. In many cases, these problems arise from the group's cooperations with industrial partners. When open application problems are abstracted into well-posed problem settings, they often provide inspiring research challenges. The group's work on spam and email security, personalized relevance ranking, and prediction of therapy outcomes falls into this category.

The research group's technology-oriented research focused on abstract, more general problem settings that are not yet understood and have no satisfactory known solution. In this line of work, the research goals are to understand properties of challenging problem settings, to extend known technologies or develop new technologies that can be shown to address these problem settings under defined circumstances. Current research on learning from differing training and test distributions, and adversarial prediction falls into this category.

Transfer Learning

Most machine learning algorithms are constructed under the assumption that the training data be governed by the exact same distribution which the model will later be exposed to.

In practice, control over the training data is often less perfect. Sample selection bias may be the result of collection protocols that rely on voluntary responses; training data may be obtained under laboratory conditions that cannot be expected after deployment of a system; image processing systems may be deployed to foreign geographic regions where vegetation and lighting conditions result in a distinct distribution of input patterns.

The Machine Learning Group has contributed several results to the area of transfer learning. The first such result is a logistic regression model for covariate shift; in this case, only the marginal input distributions differ. Given labeled training data and unlabeled test data from a different distribution, the model minimizes the expected loss on the test distribution. The second such result is a technique for multi-task learning. In this problem setting, one seeks to solve several classification problems. Some of the classification problems will likely relate to one another, but one cannot assume that the tasks share a joint conditional distribution of the class label given the input variables. The group has contributed a mathematically rigorous multi-task learning model that can handle arbitrarily different data distributions for different tasks without making assumptions about the data generation process or the relation between tasks. One application that this model can solve better than more ad-hoc models is the prediction of the outcome of HIV therapies. A cooperation of Research Group 2 and Department 3 has led to a model that can accurately predict the outcome of a therapy attempt for a specific version of the HI virus and any given candidate combination of drugs.

Finally, the group has produced a mathematically clean solution to transfer learning problems in which the input distribution is biased, and auxiliary data sources are available in which the joint distribution of input and output diverges from the target distribution. This solution has applications in personalized information ranking. The group's work on transfer learning has been distinguished with a Google Research Award.

Structural Learning

Learning mappings between arbitrary structured and interdependent input and output spaces covers many challenging learning tasks such as producing sequential or tree-structured outputs, or predicting properties of nodes in a graph. It challenges the standard model of learning a mapping from independently drawn instances to a small set of labels. Potential applications include named entity recognition and information extraction (sequential output), natural language parsing (tree-structured output), classification with a class taxonomy – here, the output is a node in a tree –, and collective classification where the output is a set of interdependent class variables.

Work by Peter Haider and Ulf Brefeld on supervised clustering has been distinguished with the Best Student Paper Award of the International Conference on Machine Learning. In this problem setting, the goal is to learn a a similarity function such that a clustering algorithm produces correct clusterings of the data. Examples of correct clusterings are provided in the training data. Ulf Brefeld, Thoralf Klein, and Tobias Scheffer have obtained several results on transductive learning and semi-supervised classification algorithms that produce sequential, tree-structured and graphical output.

Adversarial Learning

Most research on machine learning – and in fact all research in statistics – relies on the assumption that the system to be modeled does not actively resist modeling. Many security-related learning problems involve learning about an *adversary* who does attempt to foil the learner. The group has obtained the first known substantial results on Nash-equilibrial prediction models. Our results show under which circumstances a unique Nash equilibrium exists, and lead to an algorithm that learns an equilibrial prediction model from data.

Modeling the Web of Science

Publication repositories contain an abundance of information about the evolution of scientific research areas. Work of the group has addressed the problem of creating a visualization of a research area that describes the flow of topics between papers, quantifies the impact that papers have on each other, and helps to identify key contributions. To this end, we devised a probabilistic topic model that explains the generation of documents; the model incorporates the aspects of topical innovation and topical inheritance via citations. The model has been implemented in a citation influence browser that visualizes the impact that research papers have had on one another.

Learning from Streams

Many data streams have too high a volume for any knowledge discovery algorithm to process them in their entirety. The group has investigated on algorithms that learn from infinite streams efficiently and can yet be guaranteed to come ε-close to the solution that would be attained if all data were to be processed. In some interesting application scenarios, background knowledge is available in a Bayesian network. Integrating this background knowledge into the discovery process leads to algorithms that find the most surprising, unexpected patterns, rather than patterns that are already well-known.

Projects and Cooperations

A collaboration with the European web hosting company STRATO AG focuses on a range of security issues and adversarial learning. In tight collaboration with STRATO engineers, the group has contributed to the email security infrastructure of STRATO, as well as to software that detects abuse of services. Open problems at STRATO have in several cases inspired fundamental abstract research questions and have led to principled results.

In a collaboration with nugg.ad AG, the group investigates on efficient algorithms that predict user interests and demographic attributes based on the user's web browsing history.

The group has received a Google Research Award and cooperates with Google on transfer learning.

A cooperation with Fresenius-affiliate NephroCare aims at understanding the dependencies between control parameters of dialysis treatment and the long-term course of the disease. In particular, the question arises whether an analysis of the extensive data collection of Fresenius hospitals allows to optimize the treatment decisions.

Andreas Zeller of the Saarland University and Tobias Scheffer received an IBM Jazz Faculty Grant. They developed a plug-in that learns from collaboration and defect data as tracked by Jazz, relates features of the collaborative development process to the defect density of individual components, and thereby automatically predicts code quality.

In the *Text Mining* project, funded by the German Science Foundation DFG, the group studies problems related to discovering knowledge in text collections, and using this knowledge in order to improve the underlying processes.

The group has collaborated with Departments 5 (Databases and Information Systems) and 3 (Computational Biology); the latter collaboration has resulted in transfer learning techniques that allow to more accurately predict the outcome of HIV therapies.

8 The Computational Genomics and Epidemiology Group (IRG1)

Group Development

IRG1 was founded in September 2007 when Alice McHardy joined the institute. Since then four PhD students have joined the group. Dr. Ben Adams was a group member from September 2007 to February 2008, until accepting an offer from the University of Bath. The group is expected to grow slowly over the next years.

Vision and Research Strategy

Technological advances have enabled genome sequencing for organisms from a wide range of taxa, the sequencing of microbial communities from diverse environments and genome-wide surveys of genetic diversity for microbial, viral and eukaryotic populations. A major challenge is the development of computational methods for the analysis of these novel data sources. This promises to yield a more detailed understanding of biological processes with relevance for medical, biotechnological, agricultural and environmental applications.

Genomic differences between taxa or individuals of a population reflect either the impact of selection for certain phenotypic traits or neutral epidemiological processes, such as spatial population separation. Accordingly, the analysis of genomic properties such as sequence composition, gene content or individual mutations in combination with epidemiological information can give insight into the molecular characteristics and biological processes associated with phenotypic traits and allows the development of computational methods for diagnosis and classification. In terms of applications, the research of IRG1 focusses on problems in metagenomics, evolution of the influenza virus and human complex diseases. We strive to advance the state of the art for problems from each of these exciting research areas while simultaneously enhancing our expertise and skill set for solving computational genomics problems in general.

Research Areas and Achievements

In the area of metagenomics we are working on computational methods for the composition-based phylogenetic classification of metagenome sequence samples and on probabilistic models for inference of the functional context of protein families. We are actively collaborating with experimental groups from Europe, the United States and Australia, which has resulted in a Nature publication in 2007 and a Nature Biotechnology paper in 2008.

In the research of human flu, we are working on methods for vaccine strain selection for the seasonal influenza vaccine. We are furthermore investigating the role of a global reservoir of genetic diversity in the evolution and host evasion of the virus.

In the study of case-control data for human complex diseases, we are working on methods for the detection of disease-related patterns of single nucleotide polymorphisms and affected functional pathways that are associated with establishment of disease.

Projects and Cooperations

Within the institute IRG1 has established collaborations with Department 3, the former RG2 (Tobias Scheffer, Machine learning) and Department 5. Commonalities in research interests with Department 3 include the computational study of viral disease and evolution. Interactions with the former RG2 are driven by a mutual interest in adapting machine learning techniques to problems in computational genomics.

External collaborations with academia and industry

- Andreas Brune (Department of Biogeochemistry, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany)
- Ludmila Chistoserdova (Department of Chemical Engineering, Washington University, USA)
- Daniel Falush (Microbiology Department, University College Cork, Ireland)
- Jeffrey Gordon (Center for Genome Sciences, Washington University, USA)
- Phil Hugenholtz (Metagenomics Department, Joint Genome Institute, U.S. Department of Energy, USA)
- Karl-Erich Jaeger, Institut für Molekulare Enzymtechnologie, Jülich, Germany
- Mark Morrison (Metagenomics, CSIRO Livestock Industries, Queensland, Australia)
- Isidore Rigoutsos (Bioinformatics & Pattern Discovery Group, IBM T.J. Watson Research Center, USA)
- Tobias Scheffer (Machine Learning, Potsdam University, Germany)
- Andrew Weightman (Biosciences, Cardiff University, UK)

Senior Researchers

9 Molecular Networks in Medical Bioinformatics, Mario Albrecht

Overview and Research Strategy

Biological networks are at the center of life's processes. They perform many complex tasks, from the level of metabolic, regulatory, and signaling pathways of interacting molecules to that of the cell, organ, and the entire organism with its environment. An integrative strategy to gain insight into the sophisticated connections of genotype, phenotype, and environment builds from molecular biology on the smallest scale to medicine on the largest scale and requires handling diverse experimentally or computationally derived data sets. New experimental techniques for genomic, metabolic, and proteomic measurements result in large volumes of molecular data, which need to be processed, stored, combined, explored, and interpreted by efficient algorithms for biological and medical research.

To utilize this avalanche of heterogeneous data in the postgenomic era, the research group *Molecular Networks in Medical Bioinformatics* (`http://medbioinf.mpi-inf.mpg.de`) focuses on the development and application of computational methods for integrating, analyzing, modeling, and visualizing biological networks of medical relevance. Long-term objectives include integrative network models for complex diseases and the detailed elucidation and dynamic simulation of biological processes. The research group headed by Mario Albrecht was officially established within the Department of Computational Biology and Applied Algorithmics (D3) in 2006, and the group additionally became a member of the new MMCI Cluster of Excellence on the Saarland University campus (`http://www.mmci.uni-saarland.de`) in 2008 and of the local Center of Bioinformatics Saar in 2009.

The research projects cover three overlapping areas (see Chapter 29.5 for further details). The research area *Network Analysis* (Chapter 29.5.1) deals with the topological, functional, and qualitative analysis of biological networks and has a special interest in protein interactions forming molecular pathways and protein complexes functioning as molecular machines. New bioinformatics methods and databases are also implemented as user-friendly software tools or provided by web services for biological end-users. The research area *Modeling and Prediction* (Chapter 29.5.2) develops computational methods for modeling and understanding molecular networks and protein region interactions, for investigating and visualizing the effect of alternative splicing on protein interactions, and for interpreting the impact of disease-associated sequence variations on protein structure and function.

The research area *Disease Focus* (Chapter 29.5.3) aims at the development and application of bioinformatics methods for discovering and characterizing human disease genes and for advancing the systems understanding of disease processes. The current focus lies on the use of semantic similarity measures and network data for improving disease candidate prediction as well as on the application of analysis and prediction tools to viral infections (especially by

the hepatitis C virus) and polygenic auto-inflammatory diseases (an example is inflammatory bowel disease). The bioinformatic research leads to molecular hypotheses and suggests further wet-lab experiments for exploring the cause and progression of a disease as well as diagnostic and therapeutic options. In addition, it provides valuable ideas for novel and improved computational methods. Most of our methodological developments are not restricted to human cells and medicine, but also possess wide applicability in other fields like plant biology.

Research Achievements and Teaching

Chapter 29.5 provides a comprehensive summary of the group's research and publications. Most projects are organized as team work to exploit synergies and are often conducted in close collaboration with biological and medical partners. Publications within the report period include 15 conference and journal papers (cumulative ISI impact factor of 85) and one book chapter, and five master theses were completed in the group. The group has also presented numerous invited talks and posters at all major conferences, and some of the developed software tools and web sites attract many biological end-users.

Over four years, the research group has steadily grown to its current size of seven scientists (see also `http://medbioinf.mpi-inf.mpg.de/group.php`). As of April 2009, it comprises one group leader, one post-doc (Gabriele Mayr), and five PhD students (Fidel Ramírez, Andreas Schlicker, Hagen Blankenburg, Dorothea Emig, Sven-Eric Schelhorn), some of which are expected to graduate soon. Two post-docs (Christoph Welsch and Elena Zotenko) are closely involved in joint projects. In addition, the group continually hosts bachelor and master students and supervises their theses. Funding details are given in Chapter 29.5.

Recently, Dorothea Emig obtained a travel grant by the Boehringer Ingelheim Fonds (Foundation for Basic Research in Medicine) for a research visit to new cooperation partners at the Gladstone Institute of Cardiovascular Disease and at the University of California in San Francisco and Berkeley from February to May 2009. Furthermore, Mario Albrecht was a member of the scientific organizing committee of the worldwide largest bioinformatics conference ISMB/ECCB 2007 (Vienna, Austria). In May 2008, he accepted the MPG offer of a W2 group leader position in the context of the local MMCI Cluster of Excellence and also received the offer of a W3 professorship for bioinformatics at the University of Kiel.

In the last two years, the group also contributed to bachelor and master courses of protein bioinformatics. Mario Albrecht taught the lecture *Biological Networks: Databases and Analysis Methods* at the Saarland University in the summer semesters 2007 and 2008, assisted by different group members as tutors. In 2009, he gave the lecture *Molecular Networks in Biology and Medicine* at the University of Padova in Italy.

Cooperation Partners and Joint Projects

Since the research group has a strong interest in solving real biological and medical problems, networking with partners from biology and medicine is very important. This is reflected in manifold collaborations, both within the institute and external to it. A list of external cooperation partners involved with joint projects during the reporting period can be found on page 24 within the Overview chapter of the department D3.

10 Combinatorial Optimization, Ernst Althaus

Group Development

The group members were Ernst Althaus, Stefan Canzar, Daniel Dumitriu, Andreas Karrenbauer, and Rouven Naujoks. Stefan Canzar, Rouven Naujoks were PhD-students starting in 2004, Daniel Dumitriu started his PhD studies in 2007 and Andreas Karrenbauer was a post-doctoral researcher.

The group dissolved shortly after Ernst Althaus accepted a professorship at the Johannes Gutenberg-Universität Mainz in October 2008. Stefan Canzar and Rouven Naujoks finished their PhDs in November and December 2008. Stefan Canzar moved to CWI Amsterdam in February 2009. Rouven Naujoks and Daniel Dumitriu joined Ernst Althaus to the Johannes Gutenberg-Universität Mainz in March 2009. Andreas Karrenbauer moved to the group of Friedrich Eisenbrand at the EPFL Lausanne, Switzerland in October 2008.

Vision and Research Strategy

We are developing efficient algorithms for non-trivial combinatorial optimization problems which arise in different application areas.

We are considering the complete algorithm-engineering cycle: Given a vague description of a problem from a cooperation partner, we first try to agree on an abstraction as a mathematical optimization problem. Then we derive algorithms for the problems and implement them efficiently. Then we experimentally evaluate the algorithms with our partners. Ideally, they are satisfied with our solutions. Otherwise we have to revise the mathematical abstraction and work on new algorithms and implementations.

As a prototypical example where we consider the complete cycle, we refer to the project "Computing H/D exchange rates from mass-spectrometry data" (see Section 28.8.2: Interval Coloring Problem). Solution-phase hydrogen/deuterium exchange (HDX) with high-resolution mass analysis permits identification of the solvent access and contact surfaces in a protein/protein complex. Measuring HDX via mass spectrometry and analyzing the data with state-of-the-art methods reveals cumulative exchange rates of many overlapping fragments of the proteins obtained by digestion. From these rates, it is possible to infer exchange rates of single amino-acids or small parts of the protein. We provide for the first time automatic methods for solving this problem, i.e. we give an algorithm to obtain exchange rates with a higher resolution than the level of digested fragments.

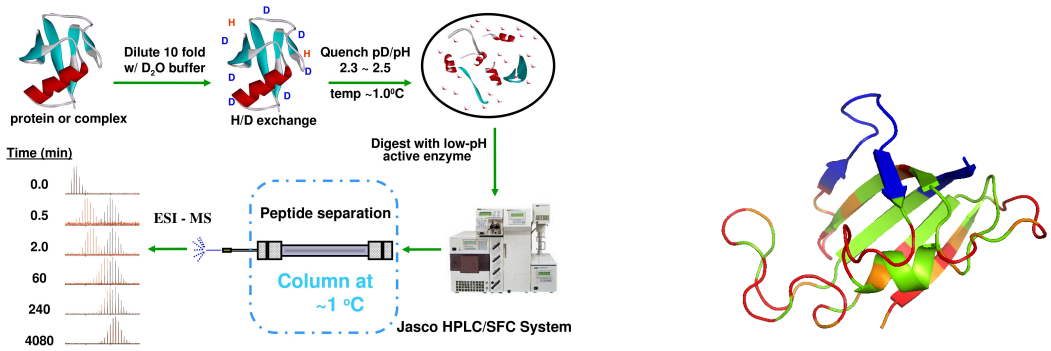


Figure 10.1: On the left, we show the schematic of HDX experiment. All steps in the HDX experiment are automated and performed by a CTC PAL robot: sample dilution, mixing, quench/digestion, timing and HPLC injection, temperature and pH are regulated throughout the experiment. On the right, we show our prediction. High exchange rates are shown in red, medium in orange and slow in green. Amino-acids for which we can't predict exchange rates as they were not covered by peptic fragments are shown in blue. One clearly observes that exchange rates are in general high on sides that are solvent accessible.

Research Areas and Achievements

Andreas Karrenbauer has been granted a patent together with his collaborators on an invention to increase the lifetime of Passive Matrix OLED devices. The algorithm, which he developed during his PhD studies, is now incorporated in Dialog Semiconductor's SmartXtend™ technology and thereby used to drive a display produced by TDK.

Projects and Cooperations

Rouven Naujoks has a position from the Transregional Collaborative Research Center 14 AVACS, where we cooperate with different research sites in Freiburg, Oldenburg, and Saarbrücken.

Daniel Dumitriu has a position within Priority Program 1307 "Algorithm Engineering" of the Deutsche Forschungsgemeinschaft (DFG), where we cooperate with PD Dr. Christoph Buchheim from the Universität Köln.

In the project "Computing H/D exchange rates from mass-spectrometry data", we are cooperating with Prof. Dr. Alan Marshall and Mark Emmett from the National High Magnetic Field Laboratory, and Prof. Dr. Anke Meyer-Bäse from the Florida State University, both in Tallahassee, Florida.

Andreas Karrenbauer is cooperating with Prof. Dr.-Ing. Xu (head of the Chair of Microelectronics at Saarland University) to analyze and solve challenging problems arising in the design of controllers for flat panel displays. In this cooperation combinatorial matrix decomposition problems are analyzed and highly efficient real-time approximation algorithms are developed.

11 Computational Chemical Biology, Iris Antes

Group Development

The Computational Chemical Biology group started as part of the Computational Chemistry group. The group consisted originally of one PhD student (Dipl.-Inf. Christoph Hartmann) and several master and bachelor students. Recently, two more PhD students joined the group (Dipl.-Biol. Kirsten Rump and Dipl.-Inf. Matthias Dietzen), who stayed at the Institute after my departure to the TU Munich and who are still co-supervised by me together with Prof. Dr. Thomas Lengauer.

Research Areas and Achievements

The goal of the group's research is to understand the dynamics and specificity of molecular recognition. For this purpose we develop novel computational methods which combine approaches from computational biology and chemistry, biophysics, and statistics. This work is performed in the context of computer-aided drug design, protein engineering, and computational immunology.

The development of highly-active drugs for an efficient and safe treatment of diseases is the major goal of pharmaceutical research. Identifying new candidates for drug design is a very challenging process during which computational and experimental scientists work hand in hand. Computational methods can aid the drug design process on several levels depending on the experimental data available. We develop new approaches for this purpose and apply these methods in specific drug design projects, which are performed in collaboration with experimental groups. Our methodological work focuses on the efficient treatment of protein flexibility during docking, and docking in situations for which an experimental structure of the receptor is not available and a theoretical structure obtained by homology modeling must be used. In molecular modeling the sampling/docking algorithm used and the corresponding scoring function are closely related, thus the development of new sampling algorithms must be accompanied by the design of the appropriate scoring functions. For this purpose we develop new scoring functions as well as software for the automated design and optimization of these functions. This work led to several new algorithms (POEM, IRECS, DynaDock) and a new software package (DynaCell).

Next to the research in computer-aided drug design we are focusing on the areas of protein engineering and structural immunoinformatics. In this context, we introduced a new method for a combined sequence and structure based prediction of peptide conformations binding to MHC class I receptors (DynaPred) and studied the effect of mutations on the function and stability of different proteins and their complexes.

Projects and Cooperations

Next to the methodological work various application studies were performed in close collaboration with experimental groups. The design and synthesis of selective inhibitors for aldosterone synthase, CYP11B2, was studied in collaboration with the Department of Pharmaceutical Chemistry (Prof. Dr. Rolf Hartmann) of the Saarland University. Homology model based docking studies were performed for the protein engineering of cytochrome P450 systems together with the Department of Biochemistry (Prof. Dr. Rita Bernhardt) of the Saarland University. Protein binding site mutations were studied for the improvement of the fungal enzyme pyranose 2-oxidase in the context of protein engineering together with Department of Applied Microbiology (Prof. Dr. Friedrich Giffhorn, Dr. Dorothee Heckmann-Pohl) of the Saarland University. Docking, molecular dynamics, and side chain prediction studies were performed for the analysis of pathogenic mutations in various protein and of drug resistance mutations in HCV protease in collaboration with Department of Internal Medicine (Prof. Dr. Zeuzem, Dr. Christoph Welsch) of the Saarland University Hospital. Structural analyses are performed for the design of epitope-specific MHC-TCR interactions in the context of adoptive T cell therapy is performed together with the Institute of Molecular Immunology (Dr. Angela Krackhardt), German Research Center for Environmental Health in Munich.

12 Information Retrieval, Hannah Bast

Group Development

The development of my group over the period of this report (April 2007 – April 2009) is characterized by the following events:

1. We have received several major science awards for our work started about four years ago, and continued in the period of this report. Details are given in the last section of this chapter.
2. In April 2008, I started a sabbatical year at Google Zurich.
3. In the summer of 2008 I received offers for tenured full-professorships from the University of Mainz (W3) and the University of Freiburg (W3).
4. Two of my students (Debapriyo Majumdar and Ingmar Weber) successfully completed their PhD at the end of 2007. Due to my forthcoming sabbatical, and at some point knowing that I would leave MPI-INF soon after that, I took on only one new PhD student (Marjan Celikik) since then.

Vision and Research Strategy

The general theme of my group is to work on algorithmically challenging problems of great practical relevance. In the past years and until this day, the focus has been on information retrieval topics, in particular, on the efficient realization of advanced search capabilities.

It is a hallmark of our work that it covers this whole spectrum / development cycle: identifying current practically relevant problems, mathematical modelling and analysis, implementation, experimentation, building of prototype system including user interfaces which deserve their name.

I consider it one of the keys of the success of our work in the past years that my role has not just been one of hiring and managing (though these are certainly very critical tasks, too), but that I was also actively and deeply involved in all parts of the mentioned development cycle myself, from proving theorems to implementing a complex Web 2.0-style user-interface with AJAX.

It should be mentioned that this working style has a price: (1) my group has always been rather small compared to the groups of other senior researchers; and (2) we do not publish in large quantity. We do, however, aim for (few, yet comprehensive) publications in top-notch conferences and journals.

Research Areas and Achievements

The focus of our research in the last four years has been on information retrieval topics, in particular, advanced search features and their efficient realization.

Examples of search features we have worked on are: auto-completion / search-as-you-type, faceted search, synonym search, search in semi-structured (XML) data, error-tolerant search, semantic search.

We have also worked on various associated machine-learning problems. (1) automatic detection of word relations (e.g.: an apple is a fruit) apple via algebraic methods purely based on word co-occurrence in the given (very large) collection, without any outside world knowledge whatsoever. (2) Efficient entity annotation, e.g. learn that the word "john" in a sentence refers to the Beatle "John Lennon", on very large text collections. (3) Find clusters of spelling variants of a word, e.g. that probalistic is a misspelling of probabilistic.

Other information retrieval topics we have worked on are: top-k query processing, advanced snippet generation, and 3D-shape retrieval (joint work with D4). We also planned to work on music information retrieval (together with Meinard Müller), but had no time for that yet.

Another big topic besides information retrieval has been route planning. We invented transit-node routing, the still fastest method for processing point-to-point shortest path queries in road networks. At Google I have worked on routing in very large public transportation networks, which, to my surprise, turned out to be an almost completely different problem.

For details on our work done during the period of this report, see the corresponding section in the part: Research Units in Detail, D1.

Projects and Cooperations

We obtained a grant from the prestigious DFG priority program "Algorithm Engineering". The title of the project is "Efficient Search in Very Large Texts, Databases, and Ontologies". The grant comprises the salary for a PhD student and for a half-time assistant (HiWi), plus some travel money. The PhD student paid from this money is Marjan Celikik. The project duration is November 2007 - November 2009.

From April 2008 – April 2009, I have spent a sabbatical year at Google Zürich. I worked on large-scale image search and route planning in very large public transportation networks. This is closely related to work I have previously done at MPI-INF, and this sabbatical has been a great asset for my scientific work in many respects. In particular, it has shown me that both the problems we work on and our working style are very much up-to-date.

Prizes and Awards

February 2007: Teaching Innovation Award, from the CS department of the University of Saarland.

October 2007: Heinz-Billing Award, together with Stefan Funke for our work on ultrafast shortest paths in road networks (5,000 Euro).

December 2007: Meyer-Struckmann Science Award, for the work on efficient search in very large databases (15,000 Euro).

October 2008: SaarLB Science Innovation Award, together with Stefan Funke for our work on ultrafast routing in road networks (25,000 Euro).

October 2008: Alcatel-Lucent Science Award, for the work on fast and intelligent search (20,000 Euro).

13 Foundations and Discrete Mathematics, Benjamin Doerr

Vision and Research Strategy

Understanding how algorithms work, or how difficult problems are, are central goals of computer science. In the research area "Foundations and Discrete Mathematics", we lay the basis for answering such questions. We investigate the mathematical structures underlying many of these problems, we develop elementary algorithms and data structures that become crucial parts of more complicated computer programs, and we analyze different complexity measures that tell us how easy or hard computational problems are.

Our main tool are mathematical proofs. Proving a statement is the ultimate assertion one can have. It also allows a degree of universality, which experimental work due to finite resources never achieves. Finally, a proof not only shows a statement, but also gives insight in why things are as they are.

The questions we try to answer are induced by computation theory itself, or they come from other areas of computer science. Some are newly asked by ourselves, others are present in the community sometimes for a long time, and again others are brought to us from researchers or practitioners in other fields.

Research Areas and Achievements

Contrary to what one would believe at first, randomized methods are highly useful in computer science. Their main application are randomized algorithms, that is, algorithms that have access to random bits ("coin flips") and may include these in their decision process. However, they are also applied in average-case analyses, or in non-explicit constructions of hard problem instances. We sketch a two areas we recently made significant progress in.

Quasirandomness. The overwhelming success of randomized methods raises the question what exactly makes them so powerful. Understanding this question allows to even further exploit these methods, both by using the advantages and by avoiding negative side-effect, which some randomized approaches have.

The paradigm of quasirandomness is the right tool to achieve these goals. It asks for analyzing the properties of a random structure and then focussing on gaining such a property, not necessarily via random approaches. Often, it turns out that it is not the randomness itself, but rather a particular consequence of using randomness, that makes a certain random structure useful.

The classical application of this concept are point sets used for numerical integration. Random point sets work reasonably well here (Monte Carlo integration). Analyzing random

point sets, we quickly note that typically they are evenly distributed in the domain of integration. Quasirandomness suggests that we try to find point sets that are particularly evenly distributed and use them for numerical integration. This led to the notion of low-discrepancy point sets, which proved to be superior for many integration problems (Quasi-Monte Carlo integration).

We were the first to use such ideas in a computer science problem. We developed a quasirandom version of the classical *randomized rumor spreading* protocol, which is used to disseminate a piece of information to all nodes of a network. For many network topologies, we could prove the equivalent or even better performance guarantees. Our experiments, however, suggest that the quasirandom protocol has a significant advantage for almost all networks.

Theoretical Analyses of Evolutionary Algorithms. Evolutionary algorithms, ant colony optimization and other bio-inspired methods are randomized search heuristics applied successfully in many practical applications. Existing research on these methods shows a strange dichotomy. On the one hand, there is huge empirical analysis supporting their extremely good performance. On the other hand, there are almost no proven performance guarantees.

A classical question, showing that even very elementary problems are not well-understood, is the following. Most evolutionary algorithms use a combination of mutation and crossover to improve a population of solution candidates. However, no problem was known where the use of crossover provably gains an advantage. Rather, obstacles like the hitchhiking effect were observed, which suggest that crossover is difficult to implement in a profitable manner.

In this light, our recent result is a great success. We were able to show (both by a mathematical proof and in experiments) that for the classical all-pairs shortest path problem the use of crossover greatly reduces the runtime of the natural evolutionary algorithm for this problem.

Subsequent work built on this result also tells us how crossover and mutation work together. Roughly speaking, crossover has not a very good chance to find a particular solution. However, it quickly finds a solution that is similar to the wanted one. These minor modifications can then efficiently be done via mutation.

Projects and Cooperations

We cooperate with various research groups on the different areas we work in. The list of co-authors of our papers shows the geographical diversity of our collaboration. Particular mention deserve our project in the priority program “Algorithm Engineering” funded by the German Science Foundation (DFG) and the collaboration with the IIT Kanpur, supported by a regular exchange and visit program funded by the IIT Kanpur.

Prizes and Awards

The papers B. Doerr, F. Neumann, D. Sudholt, C. Witt, On the runtime analysis of the 1-ANT ACO algorithm, Proceedings of GECCO 2007, pages 33-40, and B. Doerr, E. Happ, C. Klein, Crossover can provably be useful in evolutionary computation, Proceedings of GECCO 2008, pages 539–546, received a best paper award at the corresponding conferences.

14 Protein Structure and Function, Francisco Domingues

Vision and Research Strategy

Proteins play a central role in the molecular biology of all living organisms. Investigating how proteins work enables us not only to better explain life's processes, but also to understand the molecular basis of different diseases and in this way to assist the design of better therapies. Proteins are the focus of my research, in particular their structure and function, as well as their role in disease. I have been participating in the development of different methods for structure and function analysis, and I have been applying these methods to medically relevant problems.

Protein structure analysis is fundamental for the investigation of molecular function as well as the investigation of the dysfunctions associated with disease processes. In this respect I have been particularly interested in structure comparison and in the analysis of structural variation, as well as in the analysis of residue interactions. I have also been involved in the analysis of functional relationships between proteins and in the prediction of protein function based on sequence and structure relationships. Binding plays a central role in molecular function, as proteins tend to perform their function by binding to other proteins and to other types of biomolecules. I have been particularly interested in the investigation of protein binding sites and interfaces while taking into account both structural and functional aspects. The methods and the knowledge obtained from these investigations has been applied to medically relevant problems, in particular in the investigation protein interaction mimicry, viral-host protein interactions and viral drug resistance, and in the impact of mutations associated with disease.

Projects and Cooperations

Currently I participate in several projects focusing on method development, some in cooperation with other teams within our group and others involving external cooperations. There are several ongoing projects focusing on structural analysis methods, in particular there are cooperations with Peter Lackner (University Salzburg) and with Gunnar Klau (Centrum Wiskunde & Informatica Amsterdam) regarding structural comparison. There is also an ongoing cooperation with Mario Albrecht's team regarding residue interactions. We are also currently developing methods for comparing protein interfaces in cooperation with Ingolf Sommer and with Jörg Rahnenführer (Technical University Dortmund). I am also assisting Ingolf Sommer in the development of function prediction approaches.

I am also currently involved in several projects that focus on medically relevant applications. The investigation of protein mimicry and viral mimicry in particular is one of the focus of my

research. I have also been investigating viral drug resistance, in particular HCV resistance to protease inhibitors in cooperation with Christoph Welsch, and I have assisted Mario Albrecht's team in the investigation of the impact of mutations associated with disease. Finally I have been cooperating with the HIV team regarding coreceptor usage and viral-host protein interactions.

15 Real-time Rendering and Representations, Elmar Eisemann

Overview and Group Development

The research group *Real-time Rendering and Representations* headed by Elmar Eisemann has been established in October 2008 in the context of the cluster of excellence MMCI.

This research group investigates the production of realistic and expressive images. Consequently, we focus on the development of algorithms and data structures for the creation and the rendering of visual content.

Our goal is to provide methods for a (semi-)automatic production and easy manipulation of geometric content, as well as approaches to depict and simulate complex scenes.

Since April 2009, Lionel Baboud (PhD) has joined the group. He has a background in alternative representations in the context of rendering and animation.

Vision and Research Strategy

Today, content creation is one of the major cost factors in the industry. This concerns games, movie productions, illustrative and architectural design, navigation, and many other areas. One observation is that creation processes are often rather cumbersome and time consuming, e.g., creating a high-quality vector representation from a 3D model can result in hours to months of work by a skilled artist, or cleaning up scanned data sets can take weeks. One of our goals is to ease the realization by shifting tedious and repetitive tasks away from the user. Another aspect is that the development of content is often performed by users who do not necessarily understand the technical implications of their creation. Tools to bring performance considerations closer to the artist are currently rare.

Once beyond the creation/acquisition stage, data often needs to be displayed. For realistic rendering, complex computations are necessary to obtain a convincing image. This problem will not be solved by the pure increase of computational power as it is somewhat counterbalanced by the increase of the input's complexity (e.g., high-resolution scanner or volume data, or detailed models). It is hence necessary to find approximations and representations that allow us to accelerate computations like display, light transfer, and collision detection.

Finally, the purpose of visuals can be manifold. For real-time applications we usually seek to produce a single image that is discarded right after display. For illustration purposes, a high quality output is needed and it should allow various transformations without introducing artifacts. These transformations can go as far as to complete appearance changes with respect to the context it is used in. We seek to exploit and take into account these various purposes in our algorithms and to obtain an understanding of how to interact, display, modify, and represent complex data.

Research Areas and Achievements

Figure 15.1: Various Styles/Models

We accomplished several projects that represent very different aspects of our strive to improve content creation. We provide crucial, but difficult to achieve information and avoid repetitive interaction.

We developed an algorithm to evaluate visibility relationships in a scene at interactive rates. Visibility has a significant influence on the time necessary to produce images of a scene. Having this information at hand allows for the design of cost effective scenes. For example, it is possible to place simple objects in the scene which solely serve as blockers to occlude the view on more complex geometry.

In a second project, we developed a solution to convert 3D models automatically into 2D filled-region vector-art commonly found in clip-art libraries. We provided a programmable stylization to simplify, smooth and abstract filled regions. This enables us to immediately transfer the artistic appearance from one model onto the next.

The second area of investigation concerns interaction and rendering. We have shown the versatility of voxel representations in various contexts, e.g., collision detection, csg-operations, light transport. The disadvantage of voxels is that they can consume much memory. Thus, we presented a system based on several new algorithms that can render immense data volumes exceeding by far the memory capacities of the machine and representing billions of voxels, hence, Gigabytes of data.

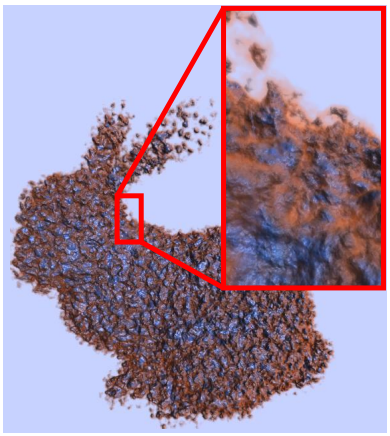


Figure 15.2: Billions of Voxels

Compared to triangular representations, voxels allow for a significant amount of details and have a better filtering behavior. This is an important point as different representations are suitable for different tasks. One general goal is to investigate ways to represent, but also convert data efficiently according to the task under which it is considered. The same holds for techniques. In this example, we combined ray-tracing and rasterization. It well illustrates the aim to unify various representations and techniques on a theoretical and practical level.

Projects and Cooperations

The group maintains a strong collaboration with researchers at INRIA (France) and initiated projects with Prof. Dr. John C. Hart (UIUC), Prof. Dr. Frédo Durand (MIT) and Adobe System Inc.

Support comes from two industrial partners: Adobe System Inc., with whom we established a longer-term partnership, and who provided us with a software grant, as well as NVidia who made their hardware available to us before its official release.

16 Combinatiorial Optimization, Khaled Elbassioni

Vision and Research Strategy

We investigate some basic problems, mainly from the computational economics and computational geometry areas, with the goal of either finding approximation algorithms with improved approximation guarantees, or ones with simpler analysis and/or description, which can translate to more efficient implementations. For geometric problems, we try to exploit geometry to develop improved approximation algorithms for problems that could be much harder in the general setting.

A common approach for developing good approximation algorithms is to model the problem as an integer program and then try to get efficiently a good approximation of its optimal solution, for instance, by considering the linear programming relaxations. At the heart of such techniques comes the study of convex polyhedra as a very central subject. We consider important properties of polyhedra that can be useful for developing such algorithms, and study the complexity of testing some of these properties.

Research Areas and Achievements

We can generally put our work into the following three different research areas.

Approximation algorithms for geometric optimization problems. We considered some basic NP-hard optimization problems arising in geometry and developed improved and/or simpler approximation algorithms for several interesting variants. In the *Traveling Salesman with neighborhoods Problem*, a salesman wants to meet a set of potential buyers. Each buyer indicates a set of potential locations where he or she can meet the buyer. The salesman would like to minimize the total length of the tour required to meet all the potential buyers. The question is how to construct such a tour. With no restriction, this problem is generally very hard. However, when the potential buyers live in 2 or 3-dimensional Euclidean space, and specify, say, convex regions as their neighborhoods, we can get much better performance guarantees. This was our main result in a number of papers (e.g. ISAAC'06, IJCGA'09), in which we improved significantly on the previously known bounds.

As another interesting example, which is a special case of *geometric set cover*, we considered the following problem: given a polygonal region in the plane determined by a monotone chain, the objective is to cover (or guard) the region with the minimum number of points (guards) from the chain. In a paper that was be presented at STACS'2009, we gave the currently best approximation factor for the problem. Unlike most of the previous techniques, our method is based on rounding the linear programming relaxation of the corresponding covering problem. Besides the simplicity of the analysis, which mainly relies on the integrality of a certain polyhedron, our algorithm also generalizes to the weighted and partial versions of the basic

problem. Currently we are working on other variants of the problem, such as guarding with demands, or maximum coverage (see Section 28.8.1 for more details).

Approximation algorithms for problems in Economics. In two papers that appeared in ESA'07 and SODA'09, we have considered the problem of *item pricing*. In the general *single-minded* setting, each customer declares a subset (bundle) of items she/he is interested in buying, with a budget (valuation) which is the maximum price she/he is willing to pay for each subset. For instance, when items are links in a network and bundles are paths to be bought or rented by different clients, the problem is known as the *tollbooth problem*. From the seller's perspective, the question is how to set the prices of the individual items such that the total profit is maximized. The situation is further complicated by the fact that there is usually a *limited supply* of each item, and sometimes also by the need to have a pricing mechanism with some *fairness* properties with respect to the clients. Recently, there has been a growing interest in this type of problems. Our main contribution here was to develop new techniques for obtaining positive and negative results, and apply them to basic variants of the problem. These techniques also opened a direction for solving other seemingly unrelated problems, such as the *constrained interval coloring problem* (SWAT'08), arising in a Bioinformatics application, and the *maximum feasible subsystems problem* (SODA'09), arising in Math programming. Currently we are working on improving the upper and lower bounds, and also extending the results to the non single-minded case.

Testing properties of polyhedra and related problems. Certain properties of polyhedra can be exploited to develop good approximation algorithms. For example, when the polyhedron corresponding to the linear programming relaxation of the problem is 0/1, i.e., each of its vertices (or extreme points) has 0/1-components, the optimal solution can be obtained by solving a linear program. Similarly, the property of being half-integral, i.e., each vertex of the polyhedron has components from $\{0, 1, \frac{1}{2}\}$ implies a 2-approximation algorithm for covering problems. In a recent paper, we have shown that testing if a given polyhedron satisfies such properties is generally NP-hard. We also considered some related problems, such as enumerating the vertices of a polyhedron or a polytope given by its description as a system of inequalities, and the *monotone Boolean dualization* problem, which is equivalent to determining if a given pair of a monotone CNF and a monotone DNF represent the same Boolean function. In a serious of papers (e.g., SoCG'08, IWPEC'08, ICALP'08) we have developed/analyzed different algorithms, and obtained new results which shed new light on the complexity of these problems (see Section 28.8.4 for more details).

Projects and Cooperations

We have cooperated with different research groups from outside the institute on different problems. In particular, our results on the TSP problem with neighborhoods and pricing problems mentioned above were obtained in joint work with Tené Sitters from the VU University, Amsterdam. Our work on the terrain guarding problem was done jointly with Domagoj Matijević from J. J. Strossmayer University of Osijek. Our work on polyhedra and Enumeration was done in cooperation with Hans Raj Tiwary from Universität res Saarlandes, Vladimir Gurvich and Endre Boros from Rutgers University, and Kazuhisa Makino from the University of Tokyo.

17 General Appearance Acquisition and Computational Photography, Hendrik Lensch

Group Development

The group on *General Appearance Acquisition and Computational Photography* headed by Hendrik Lensch has been established in May 2006. Until the end of 2008 the group grew to a size of six PhD students. Over the years the following PhD students worked in the group: Boris Ajdin, Tongbo Chen, Christian Fuchs, Martin Fuchs, Miguel Granados, Andrei Lintu, and Matthias Hullin. Three of them earned their PhD in the meantime (Andrei Lintu, Tongbo Chen, Martin Fuchs), and Christian Fuchs just handed in his dissertation. In January 2009, Hendrik Lensch accepted a professor position and moved to Ulm University.

Vision and Research Strategy

One central problem in computer graphics is the synthesis of realistic images that are indistinguishable from real photographs. The basic theory behind rendering such images has been known for a while and has been turned into a broad range of rendering algorithms ranging from slow but physically accurate frameworks to hardware-accelerated, real-time applications that make a lot of simplifications. One fundamental building block to these algorithms is the simulation of the interaction between incident illumination and the reflective properties of the scene. The limiting factor in photo-realistic image synthesis today is not the rendering per se but rather the input data passed to the algorithms. The realism of the outcome depends largely on the quality of the scene and material description. Accurate input is required for geometry, illumination and reflective properties. An efficient way to obtain realistic models is through measurement of scene attributes from real-world objects by inverse rendering. The attributes are estimated from real photographs by inverting the renderingprocess.

Research Areas and Achievements

The research of the group can be summarized in these three overlapping research areas:

3D Geometry Reconstruction. Traditional structured light 3D scanning systems are designed to capture the geometry for bright diffuse surfaces of moderate complexity. Shiny or translucent materials, e.g. metals or marble, or objects with high depth complexity typically corrupt the estimated 3D geometry producing noise or even wholes in the reconstructed surface. By designing novel capturing systems, specialized illumination patterns, and appropriate reconstruction algorithms we are able to capture the precise 3D geometry even of uncooperative static as well as dynamic objects.

Appearance Measurement. The research group further focuses on developing photographic techniques for measuring the scene's reflection properties. A so-called reflectance field captures the light transport within a scene such that all local and global illumination effects, highlights, shadows, interreflections, or caustics, are recorded and can be re-rendered under arbitrary illumination. The envisioned techniques should be general enough to cope with arbitrary materials, with scenes with high depth complexity such as trees, and should allow capturing in arbitrary environments, i.e. outside a measurement laboratory.

Computational Photography. A third thread of research of this group is computational photography with the goal to develop optical systems augmented by computational procedures; by jointly designing the capturing apparatus, i.e., the optical layout of active or passive devices such as cameras, projectors, beam-splitters, etc., together with the capturing algorithm and appropriate post-processing. Such combined systems are used to increase image quality, e.g., by removing image noise or camera shake, to emphasize or extract scene features such as edges or silhouettes by optical means, or to reconstruct volumetric 3D structures from images. We plan to devise computational photography techniques for advanced optical microscopy, large scale scene acquisition, and even astronomical imaging.

Projects and Cooperations

During the last two years a set of successful cooperations have been established to internationally renowned research institutions:

- with Prof. Marc Levoy at Stanford University we have investigated how to combine confocal imaging and pattern-based descattering to improve vision and 3D reconstruction in the presence of participating media causing multiple scattering,
- with Prof. Ramesh Raskar at MIT Media Lab (previously at the Mitsubishi Electric Research Lab) we have developed a passive display for 6D reflectance fields that renders and relights a captured 3D object in the real-world illumination incident to the display device,
- with Prof. Jan Kautz at the University College in London we are currently developing a novel acquisition system to capture the appearance of fluorescent materials, i.e. capturing their angular and wavelength dependent reflection properties.

Prizes and Awards

From the beginning the group has been supported by the Max Planck Center for Visual Computing and Communication.

Furthermore, Hendrik Lensch has been awarded an Emmy Noether fellowship by the German Research Foundation (DFG - Le1341/1-1) to build up a research group in the field of acquiring, modifying and rendering reflectance fields. The grant started in September 2007 and runs for five years with an overall budget of roughly 1,2 million Euros.

In July 2008, Hendrik Lensch received an offer for a full professor position (W3) for media informatics at Ulm University. He accepted the offer and moved to Ulm in January 2009.

18 Multimedia Information Retrieval and Music Processing, Meinard Müller

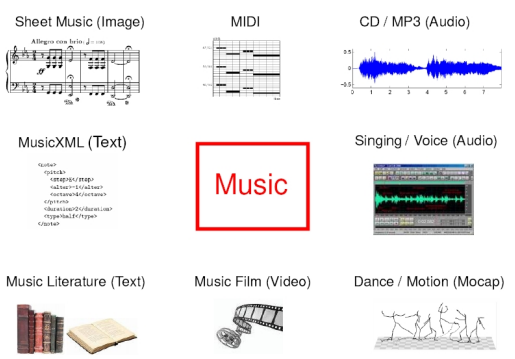
Overview and Group Development

The research group *Multimedia Information Retrieval and Music Processing* headed by Meinard Müller has been established in September 2007. Its general research mission is to advance the development of techniques and tools for analyzing, structuring, retrieving, navigating, and presenting time-dependent data streams with a focus on music data and motion data. Dealing with two different multimedia domains, the group's research falls into two major areas:

1. Music information retrieval, digital signal processing, and music processing.
2. Content-based retrieval and analysis of three-dimensional human motion capture data.

During the year 2008, four doctoral students from different disciplines have joined the research group: Dipl.-Math. Verena Konz and Dipl.-Ing. Peter Grosche are working in the music domain, whereas Dipl.-Inf. Andreas Baak and Dipl.-Inf. Thomas Helten are working in the motion domain.

Research Achievements and Vision



In the music domain, we have developed various techniques for automatic music alignment, synchronization, and matching. The common goal of these tasks is to automatically link several types of music representations, thus coordinating the multiple information sources related to a given musical work. In this context, a journal article (IEEE TASLP) as well as various proceedings articles for the main conferences on music information retrieval (ISMIR) and on acoustics, speech, and signal processing (IEEE ICASSP) were published. Prototypical implementations of user interfaces have been presented at major conferences (ECDL, ACM IMCI). As a result of digitization and the world wide web, there is an explosion of musical digital content of various types and formats comprising text, audio recordings, MIDI files, digitized sheet music, music videos, and various symbolic representations. It is the vision of the group to develop technologies to organize, understand, and search multimodal music information in a robust, efficient and intelligent manner as well as to develop tools that allow users to access, explore, and enjoy music in all its different facets.

In the motion domain, we introduced a general and unified framework for motion analysis, retrieval, and classification using binary features to represent poses. By handling spatio-temporal motion deformations already on the feature level, we are able to adopt efficient indexing methods allowing for flexible and efficient content-based retrieval even in huge motion capture databases. In this context, three articles were published (ACM MIR, JVRB, DAGM). It is the vision of the group to develop motion analysis and retrieval techniques for supporting motion classification and reconstruction from sparse and noisy sensor data.

Research Areas and Mission

Based on a rigorous mathematical foundation, the general research mission of the group is to develop generic methods that allow for efficient and deformation-tolerant retrieval of object constellations in large multimedia datasets. Important aspects concern the design of suitable features, the notion of similarity used to compare multimedia objects, as well as data organization. Here, methods are used from various fields including digital signal processing (feature extraction), computer algebra (efficiency), machine learning (analysis, classification), and information retrieval (indexing, search techniques). The group attaches great importance to dealing with research problems of practical relevance. To demonstrate the potential of the developed analysis and retrieval techniques, prototypical implementations of user interfaces and search engines have been and will be developed. Working with real-world data, many cross connections to various other research fields have been established. In the motion domain, the group is collaborating with experts from computer graphics, computer vision, and sport sciences. The music domain is interdisciplinary par excellence. Here, the group is cooperating with scientists from library science and musicology. Recently, a long-term collaboration with the University of Music, Saarland, has been established to investigate the potential of computer-based methods for music education.

Projects and Cooperations

The research group is involved in three DFG-funded research projects. The idea of the *SMART project* is to combine 3D tracking with motion retrieval techniques with the objective to stabilize markerless human motion capturing. Cooperation partners are Prof. Bodo Rosenhahn (Hannover University) and Prof. Daniel Cremers (Bonn University). In the *REKOBA project* the goal is to develop data-driven methods for reconstructing complex motion sequences from sparse and noisy sensor data in real time. In this project, the group cooperates with Prof. Andreas Weber (Bonn University). The objective of the *ARMADA project* is to develop fundamental algorithms for processing multimodal and inhomogeneous music collections with the idea of exploiting the availability of various representations. Cooperation partners are Prof. Michael Clausen (Bonn University) and Priv.-Doz. Frank Kurth (FGAN, Wachtberg). The ARMADA project member Dipl.-Inf. Sebastian Ewert at Bonn University is supervised as external doctoral student by Meinard Müller. The interdisciplinary orientation of the research group is also manifested by a recently established long-term collaboration with the University of Music, Saarland.

19 HDR Imaging and Perception Issues in Graphics, Karol Myszkowski

Group Development

The research group with the focus on the human perception aspects in rendering has been founded in June 2000 as a part of the Computer Graphics Department. Gradually the research scope has been expanded towards High Dynamic Range Imaging (HDRI), whose goal is the precise representation in appearance of real world intensity levels and color gamut at all stages of image and video processing, from acquisition to display (refer to Figure 19.1).

Within the reporting period (04.2007–03.2009) three PhD students have been successfully graduated: G. Krawczyk, K. Smith, and A. Yoshida. Also, one post-doc (and former PhD student) R. Mantiuk moved to the University of British Columbia to continue his research. At present, the group consists of one post-doc M. Okabe and three PhD students: R. Herzog, T. Aydın, and P. Didyk.

Vision and Research Strategy

The common denominator in all research efforts as conducted by the group is the advancement of knowledge on image perception and development of imaging algorithms with embedded computational models of human visual system (HVS). This way the computation performance as well as the image quality as perceived by the human observer can be significantly improved.

Figure 19.1 illustrates basic synergies between the traditional Low Dynamic Range (LDR) and HDR imaging pipelines. Our research is focused on improving the performance and image quality within HDR pipeline elements as well as securing smooth two-way transition from/to LDR devices with possible minimal information loss as perceived by the human observer. An important issue is to account for display device characteristics while designing rendering algorithms, which is often neglected in computer graphics. This becomes even more important with recently observed proliferation of display technologies, which may substantially differ in terms of reproduced contrast, brightness, screen spatial extent, pixel resolution, and dynamic response. We actively address those issues by developing tone mapping algorithms and image quality metrics sensitive to image spatial content and observer adaptation states.

Research Areas and Achievements

Tone mapping is required to accommodate HDR content to LDR devices, and conversely LDR content upgrading (the so-called inverse tone mapping) is necessary for displaying on HDR devices. In this context we addressed the problem brightness and color reproduction in tone mapped images, and we attempted to build a generic framework which generalizes

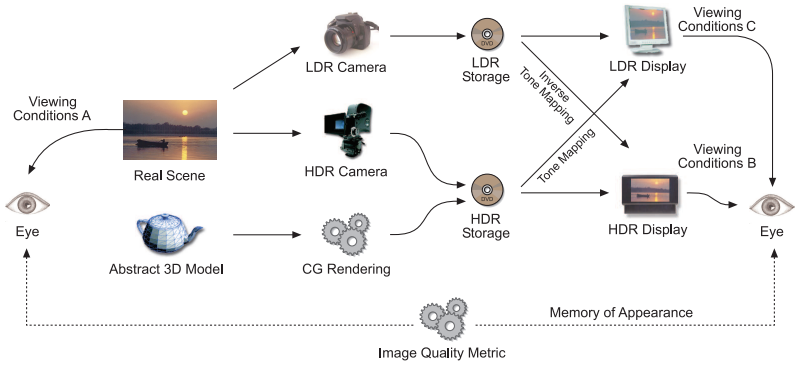


Figure 19.1: Imaging pipeline and available HDR technologies.

over image processing as performed by existing operators. Apart from considering technical capabilities of display devices, the viewing conditions such as ambient lighting and amount of light reflected by the display play an important role for proper determination of tone mapping parameters. We address this problem through optimizing over visibility of image details in tone mapped images.

Quality metrics are employed to verify algorithms at all stages of the pipeline. We have developed a quality metric, where images of different brightness and contrast ranges can be compared in terms structural changes and information loss as perceived by the human observer. We also consider state of maladaptation to account for visibility of critical information as displayed in the car and airplane cockpits in drastically different lighting conditions.

Through public releases of our tone mapping software `http://pfstools.sourceforge.net/` and image quality evaluation services `http://drim.mpi-inf.mpg.de/generator.php` significant popularity of these tools has been gained in the professional imaging community, amateur photographers and Flickr users (refer to Chapter 30.12).

Another track of our research is related with apparent image contrast and brightness enhancement beyond physical limitations of display devices. It turns out that by employing the Cornsweet illusion in the context of tone mapping, color to grey transformation, or 3D rendering apparent contrast can be significantly improved in resulting images. Effectively, as it has been observed by artists in the past, skillful manipulation with countershading (halo) effects over contrast edges can improve the overall image appearance. Brightness enhancement can be achieved by glare effect rendering. To boost the brightness impression even further, we introduce temporal glare effects, which are grounded on wave optics simulation that involves dynamic components in the human eye.

Within the reporting period the group published three ACM SIGGRAPH and nine EUROGRAPHICS papers. Also, the textbook entitled “High Dynamic Range Video” has been issued by Morgan&Claypool Publishers. For the full account of all publications and more detailed research description please refer to Chapters 30.10 and 30.11.

Prizes and Awards

G. Krawczyk *et al.* and K. Smith *et al.* have been distinguished with the 2nd and 3rd Best Paper Awards at the Eurographics 2007 and 2008 conferences, respectively.

20 Markerless Motion Capture, Bodo Rosenhahn

Group Development

The research group consists of three PhD-students, Jürgen Gall, Nils Hasler and Martin Sunkel. Jürgen Gall will defend his PhD-Thesis in June/July 2009 and Nils Hasler is expected to follow soon (autumn 2009). Bodo Rosenhahn accepted the offer for a faculty position at the University of Hannover and started as full professor in September 2008.

Vision and Research Strategy

Classical motion capture (MoCap) comprises techniques for recording the movements of real objects such as humans or animals. In biomechanical settings, it is aimed at analyzing captured data to quantify the movement of body segments, e.g. for clinical studies, diagnostics of orthopaedic patients or to help athletes to understand and improve their performances. It has also grown increasingly important as a source of motion data for computer animation.

Well known and commercially available marker based tracking systems exist, e.g. those provided by Motion Analysis, Vicon or Simi. The use of markers comes along with intrinsic problems, e.g. incorrect identification of markers, tracking failures, the need for special laboratory environments and lighting conditions and the fact that people may not feel comfortable with markers attached to the body. This can lead to unnatural motion patterns.

For these (and several other) reasons our research mission is to develop approaches for marker-less motion capture. We are interested to compute the pose and joint angles from subjects in arbitrary environments (including outdoor scenes) and apply modeling techniques from computer graphics to tackle the unsolved computer vision problems, such as segmentation, registration and pose tracking. Within several projects we even extend this basic scenario to textured models, clothed persons, subjects interacting with sports gear or moving and unsynchronized cameras.

Research Areas and Achievements

The research group *Markerless Motion Capture* deals with open questions regarding modelling, tracking, understanding and analyzing human motions from video data. In different research projects a model-based approach is proposed to allow silhouette based motion capture of parameterized free-form surface meshes. The research group is dealing with different topics, including:

- Silhouette extraction (using level-set functions, coupled with 3D shape priors)

- 2D-3D pose estimation (based on free-form surface meshes)
- Correspondence estimation
- Particle filter
- Statistical learning
- Quantitative error analysis, performance evaluation on the HumanEVA benchmark
- Cloth synthesis and animation
- Texture driven tracking
- A statistical model for human pose and shape
- Restricted kinematic chains

During 2007/2008 we published 25 book chapters, conference and journal articles including IJCV (International Journal of Computer Vision, highest impact factor in computer science) and CVPR (IEEE Conference on Pattern Recognition, acceptance rate approx. 22%).

Projects and Cooperations

In 2007 we established a successful cooperation with Daimler (Sindelfingen) and applied our approaches for markerless crash-test dummy analysis. Jürgen Gall further received a three-month internship with Microsoft Cambridge. His project was recently accepted for CVPR 2009.

Prizes and Awards

Olympus Award 2007 (Olympus Foundation)
DAGM Main Prize 2007 (German Pattern Recognition Society)

21 Efficient Search in Semistructured Data Spaces, Ralf Schenkel

Group Development

The research group *Efficient Search in Semistructured Data Spaces* has been established in December 2007. The group currently consists of two doctoral students: Andreas Broschart who is working on efficiency issues in XML and text retrieval, and Tom Crecelius who is working on search in social networks.

Vision and Research Strategy

An increasing amount of today's data is available in semistructured form, i.e., they provide more structured information than plain text documents, but don't have the regular structure of a traditional, relational database. Important examples for such semistructured information include the following: (1) texts that have been annotated with semantic markup to denote grammatical structure or real-world entities, (2) social networks with their complex relationships of people and their data, often including items like images, movies or bookmarks that have been collaboratively tagged, and (3) graph-structured knowledge networks such as ontologies.

The research in this group focuses on managing, querying and analyzing large collections of such semistructured data. Our work integrates aspects of data management and information retrieval, requiring both effective retrieval models to find relevant results and efficient data structures and algorithms to quickly retrieve these results.

Research Areas and Achievements

This subsection gives a short overview of the main research areas of the group; more details can be found in Section 31.8.9 and Section 31.9. The research done in the group currently focuses on the following topics:

Efficient search in large collections of text and XML. We are aiming at providing interactive response times for queries even on extremely large collections in the order of Terabytes or even on the Web. Besides developing efficient exact top-k algorithms, we have focused on approaches that quickly determine good approximations of the final results. Here, we develop promising heuristics for budget-aware algorithms that try to find the best approximate answers within a given budget of resource cost (like disk accesses) or within a given time. In a second line of work, we exploit extensive precomputations of indexes for term pairs, trading in improved query processing time for a higher need of disk storage.

Efficient search in social networks. Social networks like Facebook and MySpace, but also community platforms such as del.icio.us, YouTube, and Flickr have become very popular to share items such as images or links, annotate them with tags, add ratings or comments to existing content, and maintain a network of *friends*. Users search for items by providing a set of tags. Such search tasks can be classified in three different categories: social search for items contributed by explicit friends, spiritual search for items provided by users with a similar interest profile, and global search for the most frequent items (currently the dominant paradigm on these platforms). We developed a context-aware scoring function which, in contrast to standard IR query models, can be tuned towards the different search tasks. Experiments show that this score is significantly more effective than a global score that does not take the querying user and her friends into account.

On the efficiency side, we developed the *ContextMerge* algorithm for efficient topk- search and ranking in social networks. It extends existing top-k threshold algorithms over inverted lists of different types with various novel techniques for the specific setting of large online communities. For leveraging the "social wisdom", the algorithm incrementally explores the space of (explicit or implicit) friends and accumulates scores of items on the fly as they are encountered, limiting the expansion to a minimal number of related users. The algorithm can efficiently compute the best matches for social and spiritual search and, by falling back to a threshold algorithm on precomputed index lists, also for global search, even in huge communities with high dynamics.

Maintaining persistent archives of highly dynamic document collections. The World Wide Web has become a key source of knowledge, but much of the data on the Web is highly ephemeral in nature. Organizations like the Internet Archive have been working towards preserving the ever changing Web. However, they have so far paid little attention to developing a global-scale infrastructure for collecting, archiving, and *performing historical analyzes* on the collected data. Our Everlast approach, a scalable *distributed framework* for next generation Web archival and *temporal text analytics over the archive*, aims to close this gap. Everlast pursues two main goals: (1) efficiently collecting and maintaining an archive of the World Wide Web with good coverage, and (2) enabling historical text analytics by efficiently processing *Time-travel Keyword Queries* to return a ranked list of relevant documents from a web snapshot in the past. We combine standard methods of crawling with human-assisted crawlers – archival plugins in browsers or proxies which capture Web pages of archival interest and publish them into the archive. The archive itself combines existing persistence solutions from the peer-to-peer community, a storage-metadata management layer required for Web archiving, and an indexing layer for processing time-travel queries.

Cooperations

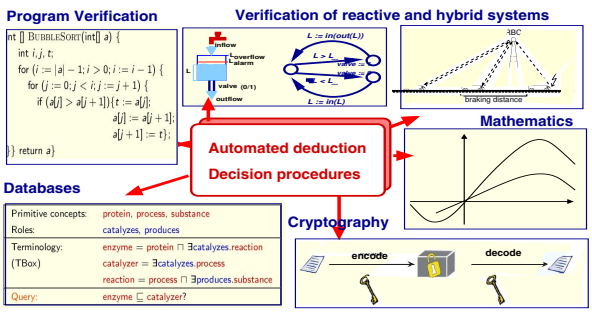
The group actively participates in the organization of the Initiative for the Evaluation of XML Retrieval (INEX): We provide the document collection for the 2009 benchmark, consisting of Wikipedia-based XML articles with semantic annotations, and Ralf Schenkel has been co-organizing the Efficiency Track since 2008 (with Martin Theobald, D5). We are collaborating with IBM in Haifa, Israel on the topic of budget-aware top-k algorithms.

22 Reasoning in Complex Theories, Viorica Sofronie-Stokkermans

Group Development

The group *Reasoning in Complex Theories* led by Viorica Sofronie-Stokkermans is part of the group *Automation of Logic*. The group consists of Viorica Sofronie-Stokkermans and two PhD students: Carsten Ihlemann and Swen Jacobs.

Vision and Research Strategy



An very important research objective in computer science is to obtain means of reasoning in and about complex systems. Systems of interest can be, e.g. programs, reactive or hybrid systems, or databases (possibly distributed; possibly changing in time). In real-life applications we may need to consider systems whose description contains all the aspects above – e.g. reactive (or hybrid) systems with embedded software, whose behavior is controlled by using databases, planning mechanisms, or programs. Verifying the correctness of such systems is extremely important – this holds especially for safety-critical systems (cars, trains, planes, or power-plants) where even small mistakes can provoke disasters. Whereas general purpose, fully automated push-button verification methods cannot exist, due to undecidability results in logic, for many concrete application domains automatic verification procedures can be designed. It is therefore very important to characterize situations in which automated verification of complex systems is possible. For this, it is essential to devise methods for:

- modeling and modular verification of complex systems;
- efficiently combining proof (or control) engines used by the component subsystems. An important step is to identify theories which are decidable, preferably with low complexity, and devise methods for (modular) reasoning in combinations of theories.

These are the main goals of our work. The type of problems we consider is motivated by applications (e.g. the verification problems we analyze in the AVACS project, but also reasoning in distributed ontologies or the verification of cryptographic protocols). The long-term goal of our research is to provide accurate automated means for the formal verification of increasingly more general complex systems. We aim at solving the verification tasks modularity, by exploiting the structure of the system at various levels (structure of the mathematical theories needed for modeling the systems, subsystem organization, communication, control).

Research Areas and Achievements

Our research focuses on the modular verification of complex systems; of particular interest is modular reasoning in complex theories. We analyze various application areas.

Modular reasoning in complex theories; decision procedures. Many problems which arise in mathematics, verification, databases and cryptography involve reasoning in combinations of theories. A large part of our research is devoted to finding methods for efficient reasoning in complex theories. We identified and studied the class of *local theory extensions*, for which efficient methods for automated reasoning exist. Our results allow to identify decidable fragments for numerous theories important in the areas mentioned above, and to provide parameterized complexity results for such theories. We implemented the method in H-PiLoT, a prover which performs hierarchical reduction for local theory extensions and delegates the proof tasks in standard theories (e.g. linear arithmetic) to other state-of-the-art SMT provers. For local theory extensions, H-PiLoT performs considerably better than these provers do when used alone: it terminates with an accurate answer (unsatisfiable/satisfiable + model) or it returns a relationship between parameters of the problems, in situations where other SMT provers return "unknown". An incremental approach (iLoRe), which often increases efficiency was also implemented. For details cf. Sect. 32.5.1–32.5.9, 32.9.5 and 32.9.6.

Modular verification of complex systems. This field of work was motivated by the necessity of understanding which properties of complex systems can be checked in a modular way. An analogy with phenomena in algebraic geometry – where links between local and global properties can be expressed in a theoretical way – allowed us to characterize a class properties which can be checked in a modular way. The results are described in Sect. 32.5.14.

Applications. We analyzed the applicability of the methods we developed on various examples from the verification of reactive, real-time and hybrid systems. We also used these methods for efficient reasoning in description logics and in mathematics and (in ongoing work) the verification of cryptographic protocols. For details cf. Sect. 32.8.3–32.8.6.

Projects and Cooperations

Viorica Sofronie-Stokkermans is a principal investigator in the Transregional Collaborative Research Center 14 AVACS (Automated Verification and Analysis of Complex Systems). Within AVACS, we have past and ongoing collaborations with the groups of Prof. Dr. Damm and Prof. Dr. Olderog (University of Oldenburg), the group of Prof. Dr. Scholl (University Freiburg), and with Dr. Andrey Rybalchenko (MPI-SWS). We also collaborate with the automated reasoning group at LORIA Nancy and at the University Milano (joint meetings on the topic of reasoning in complex theories). A cooperation with Prof. Dr. Baader (University Dresden) is planned. Since the problems we consider, and the tools we use, span a wide variety of fields (automated reasoning, symbolic computation, complexity, verification, cryptography, databases), we started building bridges between these communities, by organizing interdisciplinary workshops: ADDCT'07 and '09, CEDAR'08 (for enhancing the links between the automated reasoning and complexity communities) and "Symbolic Computation and Deduction in System Design and Verification 2008" (for establishing links between the deduction, symbolic computation and verification communities).

23 Structural Bioinformatics, Ingolf Sommer

Vision and Research Strategy

Proteins account for 15-20% of the weight of living cells (the cells also containing 70% water, 1-7% nucleic acid and various other compounds) and are responsible for performing thousands of distinct functions: for example proteins catalyze reactions, act as molecular motors, or serve as structural components of eyes, muscles, or tissue. Proteins perform these functions independently or they bind other molecules in order to perform a function.

For many proteins, structural models providing the 3-D coordinates of the atoms have been experimentally determined, or have been predicted using different computational approaches. From the 3-D coordinates it is possible to characterize and analyze the protein in terms of geometry and chemistry. Nevertheless, it is seldom trivial to infer the biological, cellular or physiological role of a protein based on structural data alone; plenty of experimental effort is often involved in analyzing the functional mechanisms of a particular protein. Protein structure databases are growing at a rapid pace. Today, in March 2009, about 52,000 experimentally resolved protein structures are in the public domain. Yet many protein structures remain without functional annotations.

My vision is to learn and understand more about the causal relation between the structures of proteins and their function by developing abstract representations and algorithms enabling us to analyze the available data.

Research Areas

In order to understand more about proteins with known structure, essentially we are following two and a half lines of work:

Structural Descriptor Methods: Structural descriptors are computational representations of protein structures. They map defined parts – and these can be functionally relevant – of protein structure into a vector space. Essential properties of descriptors are the invariance under rotations and translations of the described structures, as well as the property of mapping similar structural sites to similar representations. These representations allow for efficient comparison, and for fast retrieval of similar descriptors, thus forming a foundation to put proteins into context structurally and functionally. Furthermore, the vectorial representation is amenable to statistical learning and multivariate analysis techniques.

Analyzing and Predicting Function: Some insights can be gained by inspecting thousands of proteins at the same time. With descriptor-methods but also with other, standard, protein comparison methods proteins can be related structurally. On the other hand, functional information is available for many proteins. We link these two concepts and use them for functional analysis and for function prediction.

Using just one descriptor (or other similarity measure of proteins), a set of structurally resolved and functionally annotated reference proteins can be embedded into a high dimensional vector space. The known molecular functions annotated to the reference proteins consequently map into that space. Various regions in this space exhibit considerable difference in the local conservation of molecular function. The local conservation of function with respect to structure can be modeled and used for analysis as well as for prediction.

Applying Developed Methods to Relevant Proteins: The above-mentioned high-throughput analysis is grounded with several case studies. These arise from other projects in the department, or they are triggered by Francisco Domingues. Often the proteins involved are related to disease. In conjunction with standard bioinformatics tools, our methods can help to better understand these proteins.

24 Algorithmic Game Theory and Online Algorithms, Rob van Stee

Group Development

The research group *Algorithmic Game Theory and Online Algorithms* was established in April 2008. The main topics of research in this group are algorithmic mechanism design, the price of anarchy and price of stability for various scheduling and network design problems, and online algorithms.

Vision and Research Strategy

Many classical combinatorial optimization problems take on a new flavor when considered on the Internet. Classically, we usually assume that we are given some input (in a block, so that we then have all the data that we need), and calculate (fix) some output or assignment in a centrally controlled manner, without having to deal with outside agents. By contrast, the Internet is by its nature distributed, asynchronous, uncoordinated, and divided into many sites or areas that are controlled by agents that are principally interested in their own welfare. Therefore, if we want to carry out projects on the Internet that involve several or many such agents, we need to ensure that cooperation is in the best interest of these agents, since we cannot control their behavior deterministically. Moreover, we cannot in general be sure that the data that these agents provide us with is accurate, especially if they could somehow benefit from giving us false information.

Many new and interesting problems arise in this setting, and our group investigates such problems from two directions. The first is that of *algorithmic mechanism design*, which focuses on getting selfish agents to reveal their true (private) data in order to optimize performance. This requires using an auction mechanism or (more often) designing suitable (monotone) approximation algorithms for the problem at hand. In the context of the Internet, where users appear and disappear continuously, it is also natural and important to consider *online* algorithms for such problems.

We are also interested in directly examining the effects of selfishness in combinatorial optimization. That is, we consider Nash equilibria, which are stable states in which no agent has a motive to deviate from their strategy, and compare them to the best possible solution that could be reached in a centralized setting, where there is a single controller and no freedom for the agents. Understanding the worst-case distance of a Nash equilibrium from the social optimum in simple situations is also a prerequisite for making progress in the area of (algorithmic) mechanism design. This worst-case distance is known as the coordination ratio or the *price of anarchy*. Taking a more positive point of view, the *price of stability* is

the ratio of *best* equilibrium and the best optimal design. The best Nash equilibrium is the best solution that can be proposed from which no user will 'defect'.

The second major topic of our group is *online algorithms*. In an online problem, the input arrives incrementally and irrevocable decisions need to be made before the next part of the input arrives. This models many real-world problems where it is infeasible to wait until the entire input has arrived before making any decisions. We compare the results that can be obtained under such conditions to the best possible solutions when the entire input is given in advance. Thus the focus here is on the cost of a lack of information, in contrast to the area of approximation algorithms (to which I have also contributed), where we focus on the effect of having limited (i.e., polynomial) computation time.

Research Areas and Achievements

Algorithmic Game Theory. We consider the problem of maximizing the minimum load for machines that are controlled by selfish agents, who are only interested in maximizing their own profit. Unlike the classical load balancing problem, this problem had not been considered for selfish agents before. We consider the version where the machines are related. Hence, each agent has a single private value, which corresponds to the speed of its machine. We give monotone (and hence truthful) approximation schemes for this problem, as well as a fast monotone approximation algorithm. This algorithm has approximation ratio $\min(m, (2+\varepsilon)s_1/s_m)$ where $\varepsilon > 0$ can be chosen arbitrarily small and s_i is the (real) speed of machine i. Finally we give improved results for two machines.

We also consider scheduling on uniformly related machines in the case that jobs (instead of machines) are controlled by selfish agents, and each agent prefers a machine with small load. While previous studies either consider identical speed machines or an arbitrary number of speeds, focusing on the number of machines as a parameter, we consider the situation in which the number of different speeds is small. We reveal a linear dependence between the number of speeds and the POA. For a set of machine of at most p speeds, the POA turns out to be exactly $p + 1$. The growth of the POA for large numbers of related machines is therefore a direct result of the large number of potential speeds.

Online Algorithms. We continue the study of the online unit clustering problem, which was introduced by Chan and Zarrabi-Zadeh. We design a deterministic algorithm with a competitive ratio of 7/4 for the one-dimensional case. This is the first deterministic algorithm that beats the bound of 2. It also has a better competitive ratio than the previous randomized algorithms. Moreover, we provide the first non-trivial deterministic lower bound (1.6), improve the randomized lower bound to 1.5, and prove the first lower bounds for higher dimensions. We also study several variants and generalizations of the online unit clustering problem, which are inspired by variants of packing and scheduling problems in the literature.

Approximation Algorithms. We consider the problem of packing rectangles into bins that are unit squares, where the goal is to minimize the number of bins used. All rectangles have to be packed non-overlapping and orthogonal, i.e., axis-parallel. We have presented algorithms for this problem with an absolute worst-case ratio of 2, first for the case where items can be rotated by 90 degrees and very recently also for the general case. These results are optimal provided $P \neq NP$.

25 Integrative Scientific Computing, Robert Strzodka

Group Development

The research group *Integrative Scientific Computing* headed by Robert Strzodka within the Max Planck Center for Visual Computing and Communication (MPC-VCC) has been established in November 2007 and comprises two PhD students Zhao Dong and Mohammed Shaheen.

Vision and Research Strategy

Beside theory and experiment, computational science established itself as a third valuable mode of scientific investigation. We gain more and more insights from very large and detailed simulations of natural phenomena. But it is a long way from the physical modeling of these phenomena, resulting often in differential equations, over the numerical discretization, the algorithm design, the software implementation, down to the execution on the hardware level. Many scientist work on the optimization of the different links of this chain, but local optimizations without further going considerations do not necessarily result in an overall improvement.

Our research focuses on significant improvements of performance and accuracy in scientific computing through a global optimization across the entire spectrum of continuous modeling, numerical analysis, algorithm design, software implementation and hardware acceleration.

The concatenation of individually optimal solutions on each of these layers often performs poorly due to conflicting requirements at the interfaces. Consequently, the integration of individually suboptimal but inter-coordinated solutions from all layers can be far superior. Even when the application complexity prevents a global optimization the integrative consideration of several layers already proves to be beneficial.

Currently, we concentrate on data layout and solver design for complex domains that exploit the great performance of parallel coprocessors like GPUs, Larrabee or Cell BE without exposing their restrictions and peculiarities to the application programmer.

Research Areas and Achievements

High performance simulation code involves so many disciplines that modifications are often not evaluated for the entire system but only from a local perspective. In the following we look at several examples that demonstrate how the bridging of disciplines in the solution process leads to superior performance of the entire system.

An important goal of each computation is to achieve a certain minimum accuracy required for the application. The naive approach is simply to increase the computational precision, but from a numerical analysis point of view the uniform increase is a bad option and looking at the hardware level, doubling the number precision leads to a quadrupled rather than just doubled computational effort. An integrative approach across the disciplines is to use a *mixed precision method* that concentrates high precision only in relevant parts and benefits from faster hardware execution every else.

Because of power dissipation limits micro-processors now embrace explicit parallelism rather than frequency scaling. This requires software to change radically and poses a great problem for established software packages with hundreds of thousands of lines of code. Rather than requiring the application scientist to rewrite all of this code according to the new hardware we develop a *minimally invasive* integration strategy for parallel coprocessor, that allows to use their acceleration potential without modifying the application code.

The shift towards explicit parallelism has produced several competing high performance architectures with different low level interfaces. This widens the gap between the desired high expressiveness of languages for scientific problems and the low level programming requirements. We work on bridging this gap by offering a unified interface to the different coprocessors that enables efficient representation of complex data structures and hides the peculiarities of the different hardware.

The growing disparity between computational on-chip performance and data transfer bandwidth to the chip renders naive algorithm design that focuses solely on a minimal number of instructions quite inefficient. An integrative consideration of new algorithmic ideas that optimize for coherent memory usage and a more complex utilization of existing hardware resources are necessary. We work on new space-time traversals schemes that generate more homogeneity for the hardware on the fine level but offer more heterogeneous flexibility for the algorithm across the entire space-time domain.

In summary, we can say that many of todays integration challenges in computational science come from the radical changes to the hardware that occurred within the last five years. The very different type of computing requirements must be taken into consideration on all higher levels from the implementation and algorithm design up to numerical schemes and modeling. Otherwise, the performance gap will exacerbate exponentially as the hardware continues to scale with Moore's Law.

Projects and Cooperations

The work on mixed precision methods and the coprocessor integration into clusters is performed in collaboration with the group of Prof. Turek at the University of Dortmund and Jamaludin Mohd-Yusof and Patrick McCormick from the Los Alamos National Laboratory. Work on parallel coupling of grids and particles is developed together with the group of Prof. Kolb at the University of Siegen. The hardware interoperable representation of complex domains involves coworkers from the group of Prof. Owens from the University of California, Davis. Industrial project support for the computational usage of the parallel coprocessors comes from NVidia, AMD and Intel.

26 Image-based 3D Scene Analysis, Thorsten Thormählen

Overview and Group Development

The independent research group *Image-based 3D Scene Analysis* headed by Thorsten Thormählen was established in November 2007 and currently comprises four PhD students: Nils Hasler, Kristina Scherbaum, Sergey Kosov, and Christian Kurz.

Vision and Research Strategy

The group develops new tools and algorithms for 3D scene analysis from image sequences or video. This comprises estimations of camera motion of one or more cameras in a given scene; of static scene geometry; of motion and shape of moving objects; and of scene illumination. The tools and algorithms can be used for the generation of visual effects in movie and TV production. Specialized versions of these algorithms can also be applied in the field of 3D scene analysis for autonomous aerial, ground, and water vehicles as well as industrial, domestic, and medical robots. A particular focus is put on tools and algorithms that have the potential to be turned into mass market applications and are therefore attractive for commercial partners.

Research Projects and Achievements

In the following two selected research projects are described in more detail:
3D-Modeling by Ortho-Image Generation from Image Sequences. In this project a semi-automatic approach is developed that enables the generation of a high-quality 3D model of a static object from an image sequence that was taken by a moving, uncalibrated

Figure 26.1: Left to right: frames from the input video sequence; ortho-images that are automatically assembled out of a large number of different input frames; final model that is modeled manually in a 3D modeling package by using the ortho-images as blueprints.

consumer camera. A bounding box is placed around the object, and orthographic projections onto the sides of the bounding box are automatically generated out of the image sequence. These ortho-images can be imported as background maps in the orthographic views (e.g., the top, side, and front view) of any modeling package. Modelers can now use these ortho-images to guide their modeling by tracing the shape of the object over the ortho-images. This greatly improves the accuracy and efficiency of the manual modeling process. An additional advantage over existing semi-automatic systems is that modelers can use the modeling package that they are trained in and can thereby increase their productivity by applying the advanced modeling features the package offers. The research project resulted in a publication at the Siggraph 2008 conference.

Markerless Motion Capture with Unsynchronized Moving Cameras. In this project an approach for markerless motion capture (MoCap) of articulated objects, which are recorded with multiple unsynchronized moving cameras is developed. Instead of using fixed (and expensive) hardware synchronized cameras, this approach allows to track people with off-the-shelf handheld video cameras. To prepare a sequence for motion capture, we first reconstruct the static background and the position of each camera using Structure-from-Motion (SfM). Then the cameras are registered to each other using the reconstructed static background geometry. Camera synchronization is achieved via the audio streams recorded by the cameras in parallel. Finally, a markerless MoCap approach is applied to recover positions and joint configurations of subjects. The corresponding paper is accepted for publication at the Conference on Computer Vision and Pattern Recognition (CVPR 2009).

Figure 26.2: Tracking of a person in an outdoor scene with 4 moving cameras: pose result at one time instance overlaid into the 4 camera views (left) and reconstruction of the of the scene, including the camera motion, static scene background, and the tracked person (right).

Cooperations

Within the Cluster of Excellence *Multimodal Computing and Interaction* the conjoined research project *Computer-based Auditive and Visual Scene Analysis (CAVSA)* is carried out with the Telecommunication Lab, Saarland University, which is headed by Prof. Dr.-Ing. Thorsten Herfet. The group also participates together with Prof. Dr. Anton van den Hengel, School of Computer Science, University of Adelaide, in the DAAD Group of Eight Australia - Germany Joint Research Cooperation Scheme 2009. The conjoined project is titled *Automated large-scale 3D reconstruction from multiple-station panoramas.* Furthermore, the group maintains links to the University of Hannover and Microsoft Research Cambridge.

27 Statistical Geometry Processing Group, Michael Wand

Group Development

The *Statistical Geometry Processing Group* headed by Michael Wand has been established in September 2007. The group currently consists of three PhD students (Jens Kerber, Art Teves, and Martin Sunkel) and one Postdoc (Ruxandra Lasowski).

Vision and Research Strategy

The objective of the research group is to investigate statistical data analysis techniques for geometric data sets. Informally spoken, the long term goal of this research direction would be to device techniques to "understand" geometric data sets and use the acquired knowledge to aid geometric modeling and geometry processing. There are two main motivations for addressing these questions:

Practical relevance: There are a number of important practical problems that can be solved more efficiently using statistical data analysis: A typical application area is the processing of large 3D scanner data sets, for example "drive-by scans" of complete cities. Scanner data typically suffers from noise artifacts and large acquisition holes due to occlusion. Therefore, postprocessing with a lot of manual intervention is often necessary. Obviously, a reduction of manual labor is crucial in order to be able to deal with very large data sets. As an example, an automated pattern matching algorithm can improve the model quality: By detecting recurring pieces and repeated patterns in the data, it is possible to reduce noise and fill in holes in a plausible way fully automatically (see Figure 27.1 for a simple example).

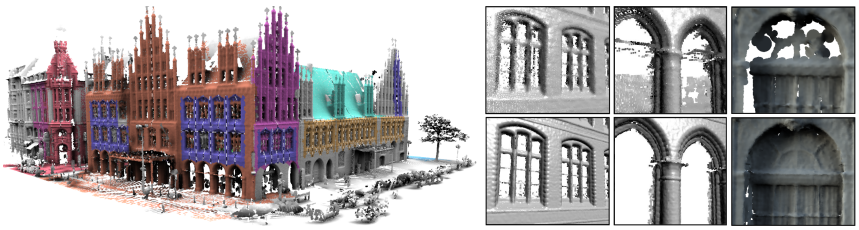


Figure 27.1: Our algorithm identifies symmetric parts. On the right: Noise removal by overlaying several symmetric parts. (Data: IKG, Leibniz University Hannover)

Philosophical insight: The second aspect that motivates this research direction is the nature of the problem itself. Before we can find algorithms that "understand the structure" of geometric data sets, we need to come up with a formal definition of what constitutes

this structure and makes it objectively identifiable. Ultimately, this leads to the question of how human perception actually works and what is necessary to define this formally. How much of what we call structure can be covered by simple mathematical formalisms and how much human knowledge (derived from millennia of cultural history) is necessary? Within the research area of statistical geometry processing, we address such problems in a forward engineering way, by devising algorithms that are capable of solving certain structure recognition problems. In this respect, our research area is very similar to the field of computer vision, which tries to "understand" images. However, by working on 3D geometry rather than image data, we avoid many of the very hard inverse problems of reconstructing information from images and thus might be able to gain new insights from this different perspective.

Research Areas and Achievements

Of course, a human level understanding of geometric data sets is very, very far from possible with currently known techniques. Therefore, we need to focus on *specific* and *solvable* problems to start with. Currently, the research group is mostly focusing on the correspondence problem, which is a key low-level problem that might serve as building block for higher level geometry analysis techniques. Establishing correspondences means that we need to find out whether two pieces of geometry are *essentially* identically, up to some noise or transformation, and compute a mapping between all points of these two pieces.

We have looked at this problem in different application scenarios. One important scenario is shape matching where parts of a shape should be assembled to a complete shape. We have developed new techniques to solve this problem not only for rigid mappings but also for deformable matching, where each piece might be deformed by a general isometric deformation. As a special case, we have applied this technique to reconstruct a single shape and its time dependent deformation from animation data from real-time 3D scanners (see Figure 27.2). A related correspondence problem scenario is symmetry detection, where correspondences within one and the same object need to be identified (Figure 27.1).

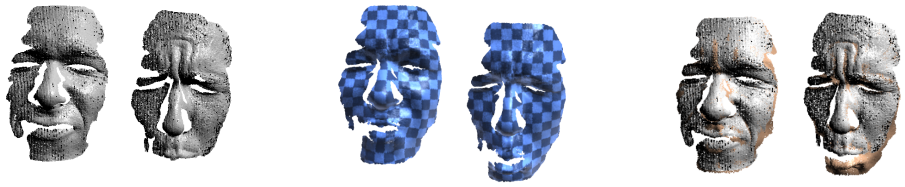


Figure 27.2: Left: Two frames of an animation. Middle: Completed shape with correspondences. Right: completion vs. single frame. (Data: Sören König, TU Dresden)

Projects and Cooperations

There is a tight collaboration with researchers from the University of Tübingen. Two additional PhD students (Alexander Berner and Martin Bokeloh) are jointly supervised with Prof. A. Schilling from the from the Tübingen computer graphics group (WSI/GRIS). In addition, there is also a collaboration with researchers from the geometry group at Stanford University (Prof. L. Guibas). In both cases, the collaborations lead to several joint publications, and is currently being actively continued in multiple further joint research projects.

Part III

Research Units in Detail

28 The Algorithms and Complexity Group (D1)

28.1 Personnel

Director
Prof. Dr. Kurt Mehlhorn

Senior Researchers and Subgroup Coordinators
Dr. Hannah Bast
Dr. Benjamin Doerr
Dr. Khaled Elbassioni
Dr. Frank Neumann
Dr. Michael Sagraloff
Dr. Rob van Stee

Researchers
Dr. Spyros Angelopoulos
Dr. Eric Berberich, on leave at Tel Aviv University
Dr. Ho-Leung Chan
Dr. Hubert Chan
Dr. Giorgos Christodoulo
Dr. Amr Elmasry
Dr. Tobias Friedrich, on leave at ICSI
Dr. Nikolaos Fountoulakis
Dr. Edda Happ
Dr. Tobias Gärtner
Dr. Michael Hemmer
Dr. Chien-Chung Huang
Dr. Andreas Karrenbauer, on leave at EPFL
Dr. Nicole Megow
Dr. Julian Mestre
Dr. Rouven Naujoks
Dr. Konstantinos Panagiotou
Dr. Rajiv Raman
Dr. Saurabh Ray
Dr. Vikram Sharma
Dr. Markus Wahlström

PhD Students
Marjan Celikik
Rolf Harren
Anna Huber
Daniel Johannsen
Kanela Kalegossi
Michael Kerber
Christian Klein
Stefan Kratsch
Ariel Rebecca Levavi
Madhusudan Manjunath
Ralf Osbild
Evangelia Pyrga
Imran Rauf
Pascal Schweitzer

Scientific Programmers
Pavel Emeliyanenko
Markus Tetzlaff
Joachim Ziegler

Secretaries
Ingrid Finkler-Paul
Christina Fries
Petra Mayer

Former (Recent) Staff
Dr. Deepak Ajwani (Postdoc, MODALGO, Aarhus)
Dr. Ernst Althaus (Professor (W2) at Universität Mainz)
Dr. Omid Amini (tenured Research Associate at Ecole Normale Superieur, Paris)
Dr. Markus Behle (Bosch GmbH)
Dr. René Beier (TomTom)
Dr. Stefan Canzar (Research Associate, CWI, Amsterdam)
Dr. Kevin Chang (Yahoo)
Dr. Majumdar Debapriyo (Microsoft, Bangalore)
Daniel Dumitriu (PhD-student, Universität Mainz)
Dr. Arno Eigenwillig (Google, Zürich)
Dr. Stefan Funke (Professor (W3), Universität Greifwald)
Dr. Joachim Giessen (Professor (W3), Universität Jena)
Dr. Anders Gidenstam
Dr. Sathish Govindarajan (Assistant Professor, Indian Institute of Science, Bangalore)
Dr. Nils Hebbinghaus (Define)
Dr. Annamaria Kovacs (Research Associate, Universität Frankfurt)
Dr. Sören Laue (Research Associate, Universität Jena)

Dr. Domagoj Matijevic (Assistant Professor, Strossmayer University, Osijek, Croatia)
Andreas Meyer
Dr. Dimitrios Michail (serving in Greek army)
Dr. Alantha Newman (Researcher, DIMACS, Rutgers)
Vitaly Osipov (PhD-student, TU Karlsruhe)
Dr. Katarzyna Paluch (Assistant Professor, University of Wroclaw, Poland)
Tobias Reithmann
Dr. Ingmar Weber (Research Associate, EPFL)

28.2 Visitors

In the time period from April 2007 to April 2009, the following researchers visited our group:

Michael Shapira	18.04.07–30.04.07	Uni Haifa
Martin Höfer	19.04.07–20.04.07	Uni Konstanz
Prahladh Harsha	21.05.07–26.05.07	TTI Chicago
Katalin Friedl	13.05.07–15.05.07	Universität Budapest
Thomas Jansen	12.06.07–13.06.07	Uni Dortmund
Günter Rudolph	12.06.07–13.06.07	Uni Dortmund
Anna Huber	11.06.07–12.06.07	Uni Kiel
Chien-Chung Huang	11.06.07–09.08.07	Sudikoff Lab. Hanover USA
Kavitha Telikepalli	02.07.07–30.07.07	Indian Inst. of Science, Bangalore
Amit Kumar	02.07.07–30.07.07	Indian Inst. Delhi
Klaus Müller	13.07.07–15.07.07	
Joerg Lehnert	19.07.07–25.07.07	Uni Frankfurt
Marcus Wahlström	01.08.07–02.08.07	Uni Stockholm
Sinan Gunturk	03.08.07–05.08.07	New York University
Petros Drineas	20.08.07–31.08.07	
Nicolò Cesa-Bianchi	03.09.07–07.09.07	Università degli Studi di Milano
Gábor Lugosi	03.09.07–07.09.07	Pompeu Fabra University, Barcelona
Hans Ulrich Simon	03.09.07–07.09.07	Ruhr-Universität Bochum
Mathias Hagen	10.09.07–14.09.07	Universität Jena
Mathew C. Francis	17.09.07–22.09.07	Indien
Michael Gnewuch	03.03.08–07.03.08	Universität Kiel
Prof. Kazhuhisa Makino	27.02.08–07.03.08	University of Tokyo
Prashant Batra	03.03.08–18.03.08	Universität Hamburg
Christine Chung	02.05.08–17.05.08	Pittsburgh University
Katrina Ligett	02.05.08–06.05.08	Pittsburgh University
Prof. Chee Yap	13.05.08.–16.05.08	University of New York
Raghav Kulkarni	15.06.08–15.08.08	
Rossano Venturini	30.04.08–30.07.08	

Kavitha Telikepalli	17.06.08–19.07.08	Indian Inst. of Science, Bangalore
Prof. Naveen Garg	23.05.08–27.06.08	IIT Delhi
Prof. Amit Kumar	19.05.08–28.06.08	
Ross Mc Connell	03.08.08–21.08.08	CMU Pittsburgh
Anupam Gupta	18.08.08–22.08.08	CMU Pittsburgh
Stefano Leonardi	18.08.08–22.08.08	Sapienza University of Rome
Seffi Naor	18.08.08–22.08.08	Technion
René Sitters	03.11.08–08.11.08	TU Berlin
Andreas Karrenbauer	10.11.08–13.11.08	EPFL Lausanne
Domagoij Matijevic	09.11.08–16.11.08	
Wiebke Höhn	11.11.08–14.11.08	
Tobias Jacobs	11.11.08–14.11.08	
Rudolf Berghammer	10.12.08–13.12.08	Uni Kiel
Jyrki Katajainen	14.01.09–27.01.09	University of Helsinki
Omid Amini	23.02.09–28.02.09	Sorbonne, Paris
Katrina Ligett	09.03.09–12.03.09	Pittsburgh University
Alejandro Lopez-Ortiz	01.02.09–19.04.09	University Waterloo, Canada
Leah Epstein	07.04.09–12.04.09	University of Haifa, Israel
Asaf Levin	07.04.09–12.04.09	Technion, Israel
Anne Auger	20.03.09	INRIA, France
Erik Jan van Leeuwen	26.03.09–28.03.09	CWI, Amsterdam
Christoph Weibel	21.03.09–25.03.09	Montreal, Canada
Ross Kang	09.03.09–11.03.09	Montreal, Canada
Dimo Brockhoff	19.03.09–20.03.09	ETH Zürich
Danny Hermelin	03.03.09–04.03.09	University of Haifa, Israel

28.3 Group Organization

The whole group meets two times per week for a scientific talk and discussions. We also have reading and discussion groups on special topics.

The whole group meets monthly to discuss organizational matters.

In the fall of every year, we have a retreat to welcome the new members of the group.

The coordinators meet biweekly.

The selection of new postdocs is made by all group members holding a PhD.

28.4 Foundations and Discrete Mathematics

Coordinator: Benjamin Doerr

In the research area "Foundations and Discrete Mathematics", we deal with basic questions and structures that are fundamental for many research activities in computer science. In spite of being fundamental, the research interests in this area do change frequently, influenced both by people joining and leaving the group and by new challenges arising from other fields.

Noting that polynomial-time solvability for many important problems is too much to ask for, we started analyzing costlier algorithms, in particular, in the framework of fixed-parameter tractability. Our previous work on graphs, one of the most fundamental structures in computer science, now has a strong focus on randomly generated graphs, which answer questions both of how a typical graph looks like and how graph algorithms typically perform.

Another type of change happened to our interest in the theory of evolutionary computation. This former sub-area, started only two years ago, grew so fast and successfully that it now forms a research area on its own, called "Bio-inspired Computation" and headed by Frank Neumann. See Section 28.5.

On the conservative side, we continue to analyze algorithms and data structures using tools from mathematics, which is a central part of today's theoretical work in computer science. The recent publication of the book "Algorithms and Data Structures: The Basis Toolbox" by Kurt Mehlhorn and Peter Sanders, a former member of the group and now a full professor in Karlsruhe, reflects the group's strong influence on this field over the last 15 years.

Our previous work on discrepancy theory lives on in our research on quasirandomness, where our quasirandom broadcasting protocol for the first time demonstrates how to use these ideas to design superior algorithms, as well as in our work on randomized rounding.

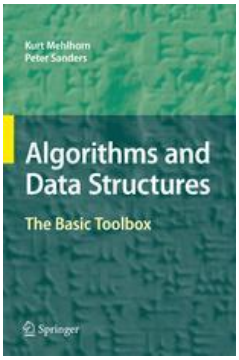
More details on these works and many others not mentioned here for reasons of space can be found on the following pages.

28.4.1 Classical Algorithms and Data Structures

Algorithms and data structures are the basic ingredients of every computer program. For this reason, they have been subject to intensive computer science research for the last fifty years at least. This does not mean that this area is now fully understood. Surprisingly, even the most classical problems, for example sorting a list of items, admit substantial advances.

Algorithms and Data Structures: The Basis Toolbox

Investigator: Kurt Mehlhorn in cooperation with Peter Sanders (Karlsruhe Institute of Technology)



Kurt Mehlhorn and Peter Sanders published a textbook "Data Structures and Algorithms: The Basic Toolbox" [1]. The book appeared with Springer Verlag and is intended for undergraduate courses in algorithms. The preface states:

"Algorithms are at the heart of every nontrivial computer application. Therefore every computer scientist and every professional programmer should know about the basic algorithmic toolbox: structures that allow efficient organization and retrieval of data, frequently used algorithms, and generic techniques for modeling, understanding, and solving algorithmic problems.

This book is a concise introduction to this basic toolbox, intended for students and professionals familiar with programming and basic mathematical language. We have used the book in undergraduate courses on algorithmics. In our graduate-level courses, we make most of the book a prerequisite, and concentrate on the starred sections and the more

advanced material. We believe that, even for undergraduates, a concise yet clear and simple presentation makes material more accessible, as long as it includes examples, pictures, informal explanations, exercises, and some linkage to the real world.

Most chapters have the same basic structure. We begin by discussing a problem as it occurs in a real-life situation. We illustrate the most important applications and then introduce simple solutions *as informally as possible and as formally as necessary* to really understand the issues at hand. When we move to more advanced and optional issues, this approach gradually leads to a more mathematical treatment, including theorems and proofs. This way, the book should work for readers with a wide range of mathematical expertise. There are also advanced sections (marked with a *) where we *recommend* that readers should skip them on first reading. Exercises provide additional examples, alternative approaches and opportunities to think about the problems. It is highly recommended to take a look at the exercises even if there is no time to solve them during the first reading. In order to be able to concentrate on ideas rather than programming details, we use pictures, words, and high-level pseudocode to explain our algorithms. A section "implementation notes" links these abstract ideas to clean, efficient implementations in real programming languages such as C++ and Java. Each chapter ends with a section on further findings that provides a glimpse at the state of the art, generalizations, and advanced solutions.

Algorithmics is a modern and active area of computer science, even at the level of the basic toolbox. We have made sure that we present algorithms in a modern way, including explicitly formulated invariants. We also discuss recent trends, such as algorithm engineering, memory hierarchies, algorithm libraries, and certifying algorithms.

We have chosen to organize most of the material by problem domain and not by solution technique. The final chapter on optimization techniques is an exception. We find that presentation by problem domain allows a more concise presentation. However, it is also important that readers and students obtain a good grasp of the available techniques. Therefore, we have structured the final chapter by techniques, and an extensive index provides cross-references between different applications of the same technique. Bold page numbers in the Index indicate the pages where concepts are defined."

References

- [1] K. Mehlhorn and P. Sanders. *Algorithms and Data Structures: The Basic Toolbox*. Springer, Berlin, 2009.

Better Heaps Supporting the decrease-key Operation

Investigator: Amr Elmasry

The pairing heap [4] is a simple self-adjusting heap that was introduced as an alternative to Fibonacci heaps [5]. Although practically more efficient [7], its theoretical analysis lags behind Fibonacci heaps. A lower bound of $\Omega(\log \log n)$ for the cost of a decrease-key operation was given by Fredman [3]. The best known upper bound is $O(2^{2\sqrt{\log \log n}})$ [6].

We gave [2] a variation of the pairing heap for which the time bounds for all the operations match the lower bounds proven by Fredman. Namely, our heap structure requires $O(1)$ for insert and find-min, $O(\log n)$ for delete-min, and $O(\log \log n)$ for decrease-key and meld.

We also gave [1] another priority queue that achieves the same amortized bounds as Fibonacci heaps. Our structure is simpler and promises a more efficient practical behavior compared to any other known Fibonacci-like heap. The main idea behind our construction is to allow for a relaxed structure by propagating rank updates instead of performing cascaded cuts after a decrease-key operation.

References

- [1] A. Elmasry. Violation heaps: A better substitute for fibonacci heaps. Research Report CoRR abs/0812.2851, CoRR, http://arxiv.org/abs/0812.2851, 2008.

- [2] A. Elmasry. Pairing heaps with O(log log n) decrease cost. In *20th ACM-SIAM Symposium on Discrete Algorithms*, New York, USA, 2009, pp. 471–476. Society for Industrial and Applied Mathematics (SIAM).

[3] M. Fredman. On the efficiency of pairing heaps and related data structures. *Journal of the ACM*, 46:473–501, 1999.

[4] M. Fredman, R. Sedgewick, D. Sleator, and R. Tarjan. The pairing heap: a new form of self-adjusting heap. *Algorithmica*, 1:111–129, 1986.

[5] M. Fredman and R. Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM*, 34:596–615, 1987.

[6] S. Pettie. Towards a final analysis of pairing heaps. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2005)*, Pittsburgh, USA, 2005, pp. 174–183. IEEE.

[7] J. Stasko and J. Vitter. Pairing heaps: experiments and analysis. *Communications of the ACM*, 30:234–249, 1987.

Improving the Comparison Complexity of Priority Queues and Deques

Investigator: Amr Elmasry

Though a basic and important data structure, the worst-case comparison complexity for priority queues was suboptimal for many years. Indeed, the upper bound for the delete-min operation for binomial queues is $2\log n + O(1)$ comparisons.

We introduced a framework for reducing the number of element comparisons performed in priority-queue operations [1]. In particular, we gave a priority queue which guarantees the worst-case cost of $O(1)$ per insert and find-min, and the worst-case cost of $O(\log n)$ with at most $\log n + O(1)$ element comparisons per delete-min.

Adding an $O(1)$-cost decrease-key to the repertoire of operations, we also gave in [3] a priority queue which guarantees at most $\log n + O(\log\log n)$ comparisons per delete-min.

In addition, we introduced two data-structural transformations to construct double-ended priority queues from priority queues [2]. With the first transformation, we obtained a deque which guarantees optimal worst-case costs and comparisons (up to constant factors), but these priority deques are not efficiently meldable. With the second transformation, we get a meldable deque having at most $\lg n + O(\lg\lg n)$ element comparisons per delete-min.

References

- [1] A. Elmasry, C. Jensen, and J. Katajainen. Multipartite priority queues. *ACM Transactions on Algorithms*, 5(1):1–19, 2008.
- [2] A. Elmasry, C. Jensen, and J. Katajainen. Two new methods for constructing double-ended priority queues from priority queues. *Computing*, 83(4):193–204, 2008.

[3] A. Elmasry, C. Jensen, and J. Katajainen. Two-tier relaxed heaps. *Acta Informatica*, 45:193–210, 2008.

Adaptive Sorting

Investigator: Amr Elmasry

A sorting algorithm is considered adaptive if it performs better for sequences having a high degree of existing order. One of the main measures of presortedness is the number of inversions in the input sequence. Many adaptive sorting algorithms have been introduced in the literature, for a survey see [2].

We studied [1] the performance of the most practical inversion-sensitive internal sorting algorithms in practice. Our objective was to conclude which of these algorithms would be a good candidate in practice, and which one we would use under different circumstances. To perform our experimentations, we selected the algorithms that we think are the most promising from the practical point of view among those introduced in the literature; these include: adaptive Heapsort, Splitsort, Greedysort, Splaysort, and adaptive AVL sort. We also compared the performance of these adaptive algorithms with randomized Quicksort.

References

- [1] A. Elmasry and A. Hammad. Inversion-sensitive sorting algorithms in practice. *ACM Journal of Experimental Algorithmics*, 13(1):1–18, 2008.

[2] V. Estivill-Castro and D. Wood. A survey of adaptive sorting algorithms. *ACM Computer Surveys*, 24:441–476, 1992.

Average-case Analysis of Dynamic Graph Algorithms

Investigators: Deepak Ajwani, Tobias Friedrich, and Nils Hebbinghaus

Many applications like pointer analysis and incremental compilation require maintaining a topological ordering of the nodes of a directed acyclic graph (DAG) under dynamic updates. All known algorithms for this problem are either only analyzed for worst-case insertion sequences or only evaluated experimentally on random DAGs. In [1] we present the first average-case analysis of incremental topological ordering algorithms. We prove an expected runtime of $O(n^2 \text{ polylog}(n))$ under insertion of the edges of a complete DAG in a random order for the algorithms of [2], [5], and [6].

In [4] we study another dynamic graph problem, namely, the fully-dynamic all-pairs shortest path problem for graphs with arbitrary non-negative edge weights. It is known for digraphs that an update of the distance matrix costs $\tilde{O}(n^{2.75})$ worst-case time [7] and $\tilde{O}(n^2)$ amortized time [3], where n is the number of vertices. We present the first average-case analysis of the

undirected problem. For a random update, [4] shows that the expected time per update is bounded by $O(n^{4/3+\varepsilon})$ for all $\varepsilon > 0$.

References

- [1] D. Ajwani and T. Friedrich. Average-case analysis of online topological ordering. In T. Tokuyama, ed., *Algorithms and Computation, 18th International Symposium, ISAAC 2007, Sendai, Japan, December 17-19, 2007. Proceedings*, Sendai, Japan, 2007, *LNCS 4835*, pp. 464–475. Springer.

[2] B. Alpern, R. Hoover, B. K. Rosen, P. F. Sweeney, and F. K. Zadeck. Incremental evaluation of computational circuits. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms (SODA '90)*, 1990, pp. 32–42.

[3] C. Demetrescu and G. F. Italiano. A new approach to dynamic all pairs shortest paths. *J. ACM*, 51(6):968–992, 2004.

- [4] T. Friedrich and N. Hebbinghaus. Average update times for fully-dynamic all-pairs shortest paths. In S.-H. Hong, H. Nagamochi, and T. Fukunaga, eds., *Proceedings of the 19th International Symposium on Algorithms and Computation (ISAAC 2008)*, Gold Coast, Australia, 2008, *LNCS 5369*, pp. 693–704. Springer.

[5] I. Katriel and H. L. Bodlaender. Online topological ordering. *ACM Trans. Algorithms*, 2(3):364–379, 2006.

[6] D. J. Pearce and P. H. J. Kelly. A dynamic topological sort algorithm for directed acyclic graphs. *J. Exp. Algorithmics*, 11:1.7, 2006.

[7] M. Thorup. Worst-case update times for fully-dynamic all-pairs shortest paths. In *Proc. of ACM Symposium on Theory of Computing (STOC '05)*, 2005, pp. 112–119. ACM Press.

28.4.2 Exponential Time Algorithms and Parameterized Complexity

Exponential Time Algorithms and Parameterized Complexity are two approaches for solving NP-hard problems, including decision problems such as Satisfiability or exact solutions to optimization problems such as Chromatic Number.

The field of Exponential Time Algorithms uses powerful algorithmic paradigms such as Branch & Reduce, Inclusion-Exclusion, and Dynamic Programming to solve some of the most difficult natural problems that computer science has to offer. Recent breakthroughs in the time complexity of some well-known problems, e.g. Chromatic Number, and in the analysis of Branch & Reduce algorithms showcase the interest in the field.

Parameterized Complexity is a two-dimensional approach for coping with NP-hardness. The central idea is to measure the time complexity of algorithms not only with respect to the size of an input instance, but also considering its structural properties. The motivation lies in the fact that reductions for proving NP-hardness create seemingly artificial instances which are unlikely to occur in practice. Moreover, it is well known that algorithms usually behave much better than a worst-case instance of a certain size would indicate. Due to this strategy some problems can be considered efficiently solved within reasonable ranges of some parameter, e.g. Vertex Cover parameterized by the cardinality of a minimum vertex cover.

Efficient Preprocessing for Generic Optimization Problems

Investigator: Stefan Kratsch

Preprocessing is a fundamental issue when one is faced with the apparent intractability of an NP-hard optimization problem. Even though polynomial-time algorithms are probably unable to compute exact solutions, it would seem ill-advised to neglect the possibility of polynomial-time preprocessing, before resorting to exponential-time techniques such as integer linear programming (which require time exponential in the input size). A major difficulty in studying preprocessing lies in quantifying its effect and in giving performance guarantees.

The field of parameterized complexity offers the notion of kernelization, allowing a rigorous study of preprocessing. Essentially, input instances to a parameterized problem come with an extra component. This so-called *parameter* is intended to express the inherent difficulty of input instances, independent of their size. A *kernelization* is a polynomial-time algorithm that reduces the size of a given instance to a value bounded by a function in the parameter, usually by means of data reduction rules. Thus the size of an instance is reduced to match its difficulty, as expressed by the parameter value. *Polynomial kernelizations* have a size bound that is polynomial in the parameter. Accordingly one can consider (polynomial) kernelizations as efficient preprocessing algorithms that come with a performance guarantee.

The last decade has seen a number of polynomial kernelizations for problems such as Vertex Cover, Feedback Vertex Set, and Triangle Packing amongst many more. Also, last year, a seminal paper due to Bodlaender et al. [1] provided polynomial lower bounds based on a hypothesis in classic (unparameterized) complexity. This implies that polynomial kernelizations are a robust way of describing preprocessing, allowing efficient algorithms as well as theoretically sound lower bounds.

We consider the kernelizability of generic optimization problems, based on a syntactic hierarchy of problem classes due to Kolaitis and Thakur [3]. For two classes we were able to present polynomial kernelizations, namely for MIN $F^{+}\Pi_1$ and MAX NP [4]. Though the definitions of the two classes are quite different, both results rely on the same tool from extremal set theory, namely the Sunflower Lemma due to Erdős and Rado [2]. Our results give an insight into the structural reason for the existence of polynomial kernelizations for a number of well-studied problems such as Vertex Cover, d-Hitting Set, Max SAT, and Max Cut. The general kernelizations that we obtain will also extend to so far unknown problems. They also give evidence for the seemingly strong connection between polynomial kernelizations and constant-factor approximability since all problems in MIN $F^{+}\Pi_1$ and MAX NP are known to admit such approximation algorithms. We feel that this underlines the importance of kernelizations as another successful strategy of coping with intractability in polynomial time.

With respect to the syntactic hierarchy of optimization problems due to Kolaitis and Thakur our results are optimal in the sense that there is no larger class, essentially allowing more alternations of quantification, to which our results could be generalized. These classes contain problems which are W[1]-hard (i.e. not even likely to admit an exponential kernelization) or problems with polynomial lower bounds using [1]. Hence, our ongoing research focuses instead on the structure of the involved constraints and formulas, to obtain a finer characterization of kernelizability.

References

[1] H. L. Bodlaender, R. G. Downey, M. R. Fellows, and D. Hermelin. On problems without polynomial kernels (extended abstract). In L. Aceto, I. Damgård, L. A. Goldberg, M. M. Halldórsson, A. Ingólfsdóttir, and I. Walukiewicz, eds., *ICALP (1)*, 2008, *LNCS 5125*, pp. 563–574. Springer.

[2] P. Erdős and R. Rado. Intersection theorems for systems of sets. *J. London Math. Soc.*, 35:85–90, 1960.

[3] P. G. Kolaitis and M. N. Thakur. Logical definability of NP optimization problems. *Inf. Comput.*, 115(2):321–353, 1994.

• [4] S. Kratsch. Polynomial kernelizations for MIN F+Pi1 and MAX NP. In S. Albers and J.-Y. Marion, eds., *26th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Freiburg, Germany, 2009. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

Counting Max-weight Solutions to 2-SAT

Investigator: Magnus Wahlström

In recent years, there has been rising interest in moderately exponential-time algorithms—that is, algorithms which solve NP-hard problems in a time that, while exponential, is much faster than what a trivial enumeration of solutions would result in. These algorithms are often some kind of branching algorithm. One experience from this research is that our tools for showing upper bounds on the running time of such algorithms are lacking. It is not only that any kind of "typical" instances would be much easier to solve than the upper bounds suggest; we know of no instances at all, even those hand-crafted to be difficult for an algorithm, that could actually exhibit a behavior approaching that of the upper bounds. An improved understanding of the behavior of these problems may lead not only to stronger bounds on the running time, but can allow for more natural algorithm constructions, as new aspects of the algorithms can be brought into the analysis, and allow for the theoretically founded use of effects that were previously considered as pure heuristics.

To some extent, such an effect has occurred with the *measure and conquer* approach to upper-bounds analysis. In 2006, Fomin, Grandoni, and Kratsch [2] presented an algorithm for the independent set problem which had not only a strong bound on the running time of $O(1.2210^n)$, but which also had a construction that was arguably more natural, with a smaller number of special cases, than many previous algorithms. Their bound is the strongest which appears in reviewed publications, but there is a technical report by Robson [3] containing a (very complicated) algorithm which, if restricted to polynomial space usage, has a bound of $O(1.2025^n)$.

In [4], we extend the toolbox by allowing a measure-and-conquer analysis to consider certain global aspects of the instance, such as the average degree, in addition to the standard accounting of local effects. The problem we consider is that of counting the number of max-weight solutions to a 2-SAT formulae, which includes the optimizing and counting variants of *independent set* as special cases. We present a simple algorithm for the problem and show a time bound of $O(1.2377^n)$, comparable to that for maximum independent set. A variant of the algorithm, with a different analysis, has previously appeared in [1].

References

[1] V. Dahllöf, P. Jonsson, and M. Wahlström. Counting models for 2-SAT and 3-SAT formulae. *Theoretical Computer Science*, 332(1-3):265–291, 2005.

[2] F. V. Fomin, F. Grandoni, and D. Kratsch. Measure and conquer: a simple $O(2^{0.288n})$ independent set algorithm. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-2006)*, 2006, pp. 18–25.

[3] J. M. Robson. Finding a maximum independent set in time $O(2^{n/4})$. Technical Report 1251-01, LaBRI, Université Bordeaux I, 2001.

• [4] M. Wahlström. A tighter bound for counting max-weight solutions to 2SAT instances. In M. Grohe and R. Niedermeier, eds., *3rd International Workshop on Parameterized and Exact Computation (IWPEC 2008)*, Victoria (BC), Canada, 2008, *LNCS 5018*, pp. 202–213. Springer.

Graph Similarity

Investigators: Kurt Mehlhorn in cooperation with N. Shervashidze, K. Borgwardt (both MPI for Biological Cybernetics), S.V.N. Vishwanathan (Purdue University), and T.H. Petri (LMU München)

What could it mean for two graphs G_1 and G_2 to be similar?

Here is a possible definition. Let k be a fixed (small) integer and let $\mathcal{G}_k$ be the set of all graphs on k vertices. We call $\mathcal{G}_k$ the graphlets of size k. $\mathcal{G}_3$ has size 4 and $\mathcal{G}_4$ has size 11, $\mathcal{G}_5$ has size 34. We associate with any graph G, the number of occurrences of each $K \in \mathcal{G}_k$ in G, i.e,

$$f_K(G) = \text{number of occurences of } K \text{ in } G, \quad K \in \mathcal{G}_k.$$

An occurrence of K in G is simply a subset S of k vertices of G such that the subgraph induced by them is isomorphic to K. We call two graphs G_1 and G_2 similar (assume for simplicity that they have the same number of vertices) if

$$f_K(G_1) = f_K(G_2) \quad \text{for all } K \in \mathcal{G}_k.$$

Two questions pose themselves. Is this a reasonable definition and can graphlet occurrences be counted efficiently?

In [1], we answer both questions in the affirmative. We show that the graphlet occurrences can be counted efficiently for graphs of bounded degree. We showed:

For $k \leq 5$ and graphs G of degree bounded by d, the vector $(f_K(G))_{K \in \mathcal{G}_k}$ can be determined in time $O(nd^{k-1})$.

We evaluated our new approach on three benchmark data sets in computational biology. These sets consist of up to 1000 graphs of size up to 1000 nodes and are hand-classified for similarity. Our method is at least competitive with existing methods. The experimental comparison showed that the discriminating power of $\mathcal{G}_k$ grows significantly with k and hence extending the theorem to larger k would be interesting. It also showed that it is important to count not only the occurrences of connected graphlets but also of disconnected graphlets.

References

- [1] N. Sherashidze, S. Vishwanathan, T. Petri, K. Mehlhorn, and K. Borgwardt. Efficient graphlet kernels for large graph comparison. In *12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Clearwater Beach, Florida USA, 2009, pp. 488–495. Society for Artificial Intelligence and Statistics.

New Plain-exponential Time Classes for Graph Homomorphism

Investigator: Magnus Wahlström

A homomorphism from a graph G to a graph H, both assumed to be simple and undirected, is a mapping $f : V(G) \to V(H)$ such that if $\{u, v\} \in E(G)$ then $\{f(u), f(v)\} \in E(H)$. The mapping does not have to be injective or surjective. The problem $\mathrm{Hom}\,(G, H)$ of deciding whether there is a homomorphism is NP-complete, and in fact, the fastest known general algorithm has a running time of $O^*(n(H)^{cn(G)})$ for a constant $0 < c < 1$. Even if the graph H is fixed, then the problem is polynomial if H is bipartite, and NP-complete otherwise [3].

Still, the exact time complexity of $\mathrm{Hom}\,(G, H)$ is an interesting open question. The problem is intermediate between two problems. On the one hand, it covers the chromatic number problem, as $\mathrm{Hom}\,(G, K_k)$ is equivalent to k-coloring, and this problem can be solved in time $O^*(2^n)$ [1]. On the other hand, it is itself a special case of 2-CSP with a domain of size $n(H)$, for which it has been shown that a $O^*(c^n)$-time algorithm for a domain-independent constant c would imply sub-exponential-time algorithms for all problems in SNP, including k-SAT [4]. Note that the trivial search space for *chromatic number* is of order n^n. The interesting question is thus whether $\mathrm{Hom}\,(G, H)$ can be solved in "plain-exponential" time, e.g., in time $O^*(2^{n(G)+n(H)})$, or whether the $n(H)^{n(G)}$-type behavior that is known is the best possible.

In [5], we consider restrictions on the graphs G and H such that the problem can be solved in plain-exponential time $O^*(c^{n(G)+n(H)})$ for some constant c. We unify and extend the previously known classes by considering graphs H of bounded *cliquewidth* [2]. A *k-expression* is an expression that creates a labelled graph using four basic operations, with k different labels used for the vertices. The cliquewidth $\mathrm{cwd}(H)$ of a graph H is the smallest k for which there is a k-expression generating H; in particular, cliques have cliquewidth 2. We show that given a k-expression for H, we can decide (or count) $\mathrm{Hom}\,(G, H)$ in time $O^*((2k+1)^{n(G)})$. To complete the result, we also show how to find a k-expression for a graph H or determine that $\mathrm{cwd}(H) > k$ in time $O^*\left((2k+1)^{n(H)}\right)$. Previously, only an approximation algorithm was known, which would return a k'-expression with $k' \leq 2^k$ if $\mathrm{cwd}(H) = k$. If H has a *core* (a homomorphically equivalent subgraph) of cliquewidth at most k, or if G has cliquewidth at most k, then we show similar results, although with a worse dependency on k.

References

[1] A. Björklund, T. Husfeldt, and M. Koivisto. Set partitioning via inclusion–exclusion. *SIAM J. Comput.*, 2006. To appear. Preliminary versions in FOCS'06.

[2] B. Courcelle and S. Olariu. Upper bounds to the clique width of graphs. *Discrete Applied Mathematics*, 101(1-3):77–114, 2000.

[3] P. Hell and J. Nešetřil. On the complexity of *H*-coloring. *J. Comb. Theory, Ser. B*, 48(1):92–110, 1990.

[4] P. Traxler. The time complexity of constraint satisfaction. In *IWPEC'08*, 2008, pp. 190–201.

• [5] M. Wahlström. New plain-exponential time classes for graph homomorphism. In *Fourth International Computer Science Symposium in Russia, CSR 2009*, Novosibirsk, Russia, 2009. Springer. To appear.

28.4.3 Discrete Mathematics

Discrete mathematics deals with mathematical objects that lack a richer structure like those investigated in algebra or topology, for example. This structural simplicity poses hard problems, however. Just the absence of structure makes it difficult to analyze such objects and derive useful properties. Since many algorithmic problems can be conveniently modelled by discrete structures like graphs, hypergraphs or simply certain functions between finite sets, there is a rich exchanges between theoretical computer science and discrete mathematics.

Enumerative Combinatorics: Counting Defective Parking Functions

Investigators: Daniel Johannsen and Pascal Schweitzer in cooperation with Peter Cameron and Thomas Prellberg

Parking functions were first introduced in connection with hashing [2]. A parking function is a function $f\colon \{1,\dots,m\} \to \{1,\dots,n\}$ with the following property. Let $f(i) = a_i$ and $a_{i_1} \leq \ldots \leq a_{i_m}$ be an increasing permutation of the function values of f, then $a_{i_j} \leq j$ must hold. There exists an explicit formula for the number of parking functions with parameters n and m.

In [1] we count the number of defective parking function. For defective parking functions the requirement $a_{i_j} \leq j$ is relaxed to $a_{i_j} \leq j + k$ for a fixed parameter k, called the *defect*. These functions can also be interpreted as cars that park in a one way street (hence the name), or in a coupon collector setting, where the coupons are of varying value and any coupon may be traded against a cheaper one.

We establish a recursive formula for the number of these functions, from which we derive a three-variable generating function. With the help of the generating function we obtain an explicit formula. The cumulative totals are partial sums that also occur in Abel's binomial formula, special cases of which can thereby be reproven. Finally we use the explicit formula to compute precise distributions of the asymptotic behavior of these functions for prominent cases. In particular, we analyze how many free parking spaces we may expect, as the parameters tend to infinity.

References

• [1] P. Cameron, D. Johannsen, T. Prellberg, and P. Schweitzer. Counting defective parking functions. *Electronic Journal of Combinatorics*, 15(1):R92, 2008.

[2] A. G. Konheim and B. Weiss. An occupancy discipline and applications. *SIAM J. Appl. Math.*, 14(6):1266–1274, 1966.

Lower Bounds for Ramsey Problems: Lovász' Local Lemma and the Incompressibility Method

Investigator: Pascal Schweitzer

The probabilistic method is a general purpose tool, used for various aspects of combinatorics. One type of application yields lower bounds on the extremal sizes of mathematical structures. A specific tool of the probabilistic method is Lovász' Local Lemma. It supplies improved bounds for problems in which the atomic events that are considered are mostly independent.

The incompressibility method is likewise used to obtain lower bounds for combinatorial problems. The underlying concept is that of Kolmogorov complexity, in particular, the incompressibility of strings, as opposed to probability spaces. The bounds obtained by a single compression step match the bounds that are obtained by an application of the method of the first moment, a basic variant of the probabilistic method. In general, when applied to probability spaces with many independent events, this yields bounds weaker than those obtained with Lovász' Local Lemma.

In [1] we show that for Ramsey type problems, (i. e., coloring problems in which a certain type of substructure is to be avoided), by using a repeated compression step within the incompressibility method, we may achieve bounds that match those obtained with Lovász' Local Lemma. Understanding the relationship between the information theoretic incompressibility method and the probabilistic Local Lemma is the first step to employ Kolmogorov complexity, in order to achieve improved lower bounds for and further insight into combinatorial problems.

References

- [1] P. Schweitzer. Using the incompressibility method to obtain local lemma results for Ramsey-type problems. *Information Processing Letters*, 109(4):229–232, 2009.

28.4.4 Random Graphs

A major task in theoretical computer science is the design and the analysis of algorithms. From today's point of view, an important indicator for the hardness of a problem is its complexity status: if, for example, we prove that a specific problem is NP-hard, then this leaves little hope for finding efficient algorithms that solve all possible instances optimally. Many significant problems appearing in several disciplines turn out to be NP-hard or even harder. However, instances of these problems appearing in practical applications seem to be mostly "easier" and are solved routinely and effectively, e.g., by heuristic approaches.

The fundamental question in this context is: Which are the *structural properties* that make an instance of a given problem "hard", and which ones make it "easy"? The knowledge and the understanding of such properties provides us with necessary information about the obstacles that efficient algorithms have to cope with. Unfortunately, the "barriers" that modern algorithms have to face are not at all well-studied.

One important step towards this direction was achieved only recently by studying constraint satisfaction problems on *random inputs*. Let us describe briefly a specific result in this context, which concerns the k-colorabilty of graphs. We denote by $G_{n,m}$ a graph drawn uniformly at random from the set of all graphs with n vertices and $m = \lceil \frac{dn}{2} \rceil$ edges. By now we have very good estimates for the largest density d for which a typical instance of $G_{n,m}$ is k-colorable.

In particular, Achlioptas and Naor [2] showed, roughly speaking, that there is a critical density d_k such that whenever $d \lesssim d_k$, then $G_{n,m}$ is typically k-colorable, while if $d \gtrsim d_k$, then the chromatic number of $G_{n,m}$ is typically greater than k. However, from an *algorithmic* point of view, despite of significant efforts over the last 35 years, all known polynomial time algorithms for coloring $G_{n,m}$ with $k \geq 3$ colors fail at densities much lower than d_k. Particularly, it was observed that there is a $d_k^* < d_k$ such that all best known polynomial-time algorithms fail if $d > d_k^*$, and succeed if $d < d_k^*$.

In a remarkable paper Achlioptas and Coja-Oghlan [1] proved that this is no coincidence, confirming a hypothesis stated by statistical physicists in 2005 [3]. They showed for several random constraint satisfaction problems that the point where the algorithm's performance breaks down is essentially the point where the *geometry* of the solution space undergoes a dramatic change. As a very rough approximation for the picture of the evolution of the solution space we mention the following: for low densities, the set of the valid k-colorings of $G_{n,m}$ looks like a giant ball. In contrast to that, at d_k^* this ball *shatters* into exponentially many small pieces that are far apart from each other. Even simple algorithms have no problem in determining valid k-colorings in the "ball" regime, but no known polynomial-time algorithm is able to find solutions in the second regime.

One positive consequence of the research mentioned above is that the study of random structures provides us with new insights about the properties of "difficult" and of "easy" instances of several computational problems. On the negative side, the classical models studied (like the $G_{n,m}$ or random k-SAT formulas) in the current literature bear properties that do not necessarily resemble characteristics of instances that appear in practical applications. A very important reason for this is that all classical models rely heavily on the concept of *independence*: a graph is generated by flipping for each possible edge a "fresh" coin, or a formula is generated by choosing each clause independently at random.

In our research we address precisely this problem. Contrary to the classical models of random structures, which are nowadays well-understood with thousands of papers devoted to their study, much less is known about typical properties of random instances from more *natural* models. We make a step towards resolving this issue by studying several different models, which have more global structure and several additional dependencies.

References

[1] D. Achlioptas and A. Coja-Oghlan. Algorithmic barriers from phase transitions. In *Proc. of 49th IEEE FOCS*, 2008.

[2] D. Achlioptas and A. Naor. The two possible values of the chromatic number of a random graph. *Annals of Mathematics*, 162(3):1333–1349, 2005.

[3] M. Mezard, T. Mora, and R. Zecchina. Clustering of solutions in the random satisfiability problem. *Physical Review Letters*, 94:197–205, 2005.

Maximal k-connected Subgraphs of Random Graphs

Investigators: Konstantinos Panagiotou and Nikolaos Fountoulakis in cooperation with Angelika Steger

In the paper [3] we started a systematic study of the properties of random graphs from classes with globally imposed structural constraints, as for example planar graphs. More specifically, let $\mathcal{C}$ be a class of labeled connected graphs, and let C_n be a graph drawn uniformly at random from graphs in $\mathcal{C}$ that contain exactly n vertices. Denote by $b(\ell; \mathsf{C}_n)$ the number of blocks (i.e. maximal biconnected subgraphs) of C_n that contain exactly ℓ vertices, and let $lb(\mathsf{C}_n)$ be the number of vertices in a largest block of C_n. We showed that under certain general assumptions on $\mathcal{C}$, C_n belongs with high probability to one of the following types:

(1) $lb(\mathsf{C}_n) \sim cn$, for some explicitly given $c = c(\mathcal{C})$, and the second largest block is of order n^α, where $1 > \alpha = \alpha(\mathcal{C})$, or

(2) $lb(\mathsf{C}_n) = \mathcal{O}(\log n)$, i.e., all blocks contain at most logarithmically many vertices.

Moreover, in both cases we showed that the quantity $b(\ell; \mathsf{C}_n)$ is concentrated for all ℓ, and we determined its expected value. As a corollary we obtained that the class of planar graphs is of Type (1). In contrast to that, outerplanar and series-parallel graphs are of Type (2).

In the subsequent work [2] we studied further the block structure of random graphs, where in addition the average degree, i.e., the number of edges, is fixed in advance. There, a similar picture as above is true.

Note that the above results give a very precise picture of random graphs from classes that are of Type (2), as they decompose into many small independent components. However, for Type (1) graphs, typically a large fraction of the vertices is "hidden" in a huge biconnected component. This motivated us to perform in [1] a closer study of the typical structure of large biconnected graphs from classes with global structural side constraints. In particular, we studied the distribution of the sizes of 3-connected components inside random biconnected graphs. To this aim, a recursive description of biconnected graphs in terms of *3-connected* graphs by Trakhtenbrot [4] turned out to be very helpful. There, a biconnected graph is decomposed into so-called *cores* of higher connectivity, i.e., 3-connected graphs. We investigated whether random biconnected graphs with respect to their 3-connected cores behave as random connected graphs with respect to the blocks they contain. Indeed, we discover that there exists also in this case a fundamental *dichotomy*: depending on a particular relation of the generating functions enumerating the 3-connected and the 2-connected graphs in the class under consideration, a random biconnected graph contains with high probability either only at most logarithmic-size cores, or a giant core that contains a constant fraction of the vertices.

All the above results give us a fairly precise description of random graphs from classes with specific structural constraints, with respect to building blocks of high connectivity. A challenging research direction in this context is to investigate how this picture can be exploited to study further the typical behavior of non-trivial parameters of such graphs. As an example, we are currently investigating the typical number of edges as well as the diameter of them.

References

[1] N. Fountoulakis and K. Panagiotou. Cores in random planar graphs. Preprint., 2009.

- [2] K. Panagiotou. Blocks in constrained random graphs with fixed average degree. In *21st International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC '09)*, Hagenberg, Austria, 2009. DMTCS (Discrete Mathematics and Theoretical Computer Science).

- [3] K. Panagiotou and A. Steger. Maximal biconnected subgraphs of random planar graphs. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '09)*, January 4-6, 2009, 2009, pp. 432–440. ACM, SIAM.

[4] B. A. Trakhtenbrot. Towards a theory of non-repeating contact schemes. *Trudi Mat. Inst. Akad. Nauk SSSR*, 51:226–269, 1958. (Russian).

Minors in Random Regular Graphs

*Investigators: Nikolaos Fountoulakis, Daniela Kühn (*University of Birmingham, U.K.*), and Deryk Osthus (*University of Birmingham, U.K.*)*

In [2], we investigate the number of vertices of the largest complete minor inside a random r-regular graph on n vertices, where $r \geq 3$ is fixed. Recall that a graph H is contained as a *minor* in a graph G, if we can obtain a copy of H by a repeated application of edge contractions, deletions and vertex deletions in G. As the number of edges of an r-regular graph on n vertices is $rn/2$, the number of vertices of any complete minor in such a graph is necessarily of order $\sqrt{n}$. In joint work with D. Kühn and D. Osthus (University of Birmingham, U.K.), we showed that the largest complete minor in a random r-regular graph on n vertices has at least $c\sqrt{n}$ vertices with probability $1 - o(1)$ as $n \to \infty$.

This result finds an application in the asymptotic estimate of the largest minor that is contained in the giant component of $G_{n,p}$. Recall that this is a random graph on n vertices where every edge appears independently with probability p. We are interested in p very close to the value $1/n$, which is the critical probability for the emergence of the giant component. This is a component whose number of vertices is proportional to n. When $p = (1 + \varepsilon)/n$, then with probability $1 - o(1)$ the largest component of $G_{n,p}$ contains cn vertices, where $c = c(\varepsilon)$, whereas every other component is of order $O(\log n)$. But when $p = (1 - \varepsilon)/n$, then with probability $1 - o(1)$ all components contain $O(\log n)$ vertices. This "phase transition" was among the early and fundamental discoveries of Erdős and Rényi [1] on the theory of random graphs. The above result on random r-regular graphs allowed us to describe the evolution as p goes away of $1/n$ of the number of vertices of the largest complete minor that is contained in the emerging giant component.

References

[1] P. Erdös and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[2] N. Fountoulakis, D. Kühn, and D. Osthus. Minors in random regular graphs. *Random Structures and Algorithms*, xx, 200? to appear.

Vertices of Degree k in Random Maps

Investigators: Daniel Johannsen and Konstantinos Panagiotou

A *map* is an embedding of a planar graph to the sphere, where generally multiple edges and loops are admissible. Maps are commonly used to describe the topology of geometric arrangements. In particular, the class of three-connected planar maps is combinatorially equivalent to the class of three-dimensional convex polyhedra.

The study of maps has a long history. Already Euler asked for the number of isomorphism types of convex polyhedra, which, by a well-known theorem of Steinitz [5], are combinatorially equivalent to three-connected planar graphs. Subsequently, Whitney [7] showed that all embeddings of such a graph are topologically equivalent, which in particular implies the existence of a simple one-to-one correspondence between three-connected planar maps and three-connected planar graphs. However, Euler's question still remains unanswered.

A general theory of map enumeration was initiated by Tutte [6] in the early 60's, who studied for the first time systematically the number of maps in various classes considered in this work. Since then, maps have been investigated extensively as combinatorial and as geometric objects, and in the meantime a rich theory highlighting several properties and aspects of maps has evolved. However, much less is known about statistical properties of random maps, that is, maps drawn uniformly at random from the class of all maps with a given number of edges. A main achievement in this context is the precise description of the so-called *core-size* of a random map, which was provided by Banderier, Flajolet, Schaeffer, and Soria [1]. Among several other results, they showed that a random map contains typically a giant (that is, linear-size) biconnected submap.

One reason for this lack of knowledge is that maps are heavily *constrained* combinatorial objects, in the sense that the appearance of specific edges is highly dependent on the presence or absence of other edges. Hence, one has to resort to exact counting techniques to obtain precise results. One of our aims is to attack precisely this problem, and to demonstrate that maps contain in a well-defined sense enough "independence", allowing us to study their typical asymptotic properties. In particular, we adapt and extend significantly an approach first used in [2], where sampling techniques were used to reduce the study of certain properties of constrained random graphs to sums of independent random variables.

Our studies concentrate on the degree sequences of several families of random maps. Based on the work of Gao and Wormald [3] on general random maps, we derive in our work [4] relations and exact asymptotic expressions for the number of vertices of degree k in random biconnected and three-connected planar maps. Moreover, we provide accompanying large deviation statements. In particular, we show that in large random biconnected maps the number of vertices of degree k is sharply concentrated around its expected value. Additionally, we demonstrate how concentration results for general random maps carry over first to biconnected and then to random three-connected maps.

References

[1] C. Banderier, P. Flajolet, G. Schaeffer, and M. Soria. Random maps, coalescing saddles, singularity analysis, and Airy phenomena. *Random Structures Algorithms*, 19(3-4):194–246, 2001.

[2] N. Bernasconi, K. Panagiotou, and A. Steger. On properties of random dissections and triangulations. In *SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, 2008, pp. 132–141. Society for Industrial and Applied Mathematics.

[3] Z. Gao and N. C. Wormald. Sharp concentration of the number of submaps in random planar triangulations. *Combinatorica*, 23(3):467–486, 2003.

[4] D. Johannsen and K. Panagiotou. Vertices of degree k in random maps. Preprint, 2009.

[5] E. Steinitz. Polyeder und Raumeinteilungen. *Enzyklopedie der mathematischen Wissenschaften*, 3(9):1–139, 1922.

[6] W. T. Tutte. A census of planar maps. *Canad. J. Math.*, 15:249–271, 1963.

[7] H. Whitney. Congruent graphs and the connectivity of graphs. *American Journal of Mathematics*, 54(1):150–168, 1932.

Synchrony and Asynchrony in Neural Networks

Investigators: Konstantinos Panagiotou in cooperation with Fabian Kuhn, Joel Spencer, and Angelika Steger

The dynamics of large networks is an important and fascinating problem. Key examples are the Internet, the world wide web, and the human brain. In particular, understanding the human brain is an as fascinating topic as it is difficult to tackle with our current level of knowledge. Almost all aspects provide more questions than answers. This reaches from understanding functionality from a high level, down to developing models for several physiological phenomena. For the latter, a topic that was intensively studied over the last decades is the dynamics of pulse-coupled oscillators of networks, which are formed by possibly leaky integrate-and-fire neurons (see eg. [7, 1, 2, 4, 6]). While all these models "live" in a continuous setup, a completely orthogonal approach was very recently taken by DeVille and Peskin [3], who replaced the deterministic continuous-state integrate-and-fire neuron by a discrete-state neural model.

The key feature and novelty in their approach is that they describe the interactions of a neuronal system as a *discrete*-state stochastic dynamical network. This idealization has two benefits: on the one hand it captures key features of neuronal behavior, and on the other hand it allows the study of spontaneous synchronization, a phenomenon in neuronal networks that is far from being well-understood.

In synchronous behavior the firing of one neuron leads to the firing of other neurons, which in turn may set off a *chain reaction* that often involves a substantial proportion of the neurons. There are strong analogies to the giant component phenomenon in random graph theory first explored by Paul Erdős and Alfred Rényi.

DeVille and Peskin show (experimentally and by some estimates in an associated mean-field limit of the model) that their network can exhibit both synchronous and asynchronous behavior. They also exhibit a range of parameters for which the network switches seemingly spontaneously between synchrony and asynchrony.

In our work [5] we rigorously analyze their model thereby answering in particular their questions about the actual parameter settings resp. thresholds for which these changes between synchronous and asynchronous behavior occur. We provide insights into the coexistence

of synchronous and asynchronous behavior and the conditions that trigger a 'spontaneous' transition from one state to another.

References

[1] P. C. Bressloff and S. Coombes. Desynchronization, mode locking, and bursting in strongly coupled integrate-and-fire oscillators. *Phys. Rev. Lett.*, 81(10):2168–2171, 1998.

[2] S. R. Campbell, D. L. L. Wang, and C. Jayaprakash. Synchrony and desynchrony in integrate-and-fire oscillators. *Neural Computation*, 11(7):1595—-1619, 1999.

[3] R. E. L. DeVille and C. S. Peskin. Synchrony and asynchrony in a fully stochastic neural network. Bulletin of Mathematical Biology, accepted for publication, 2008.

[4] P. Goel and B. Ermentrout. Synchrony, stability, and firing patterns in pulse–coupled oscillators. *Physica D: Nonlinear Phenomena*, 163(3–4):191–216, 2002.

[5] F. Kuhn, K. Panagiotou, J. Spencer, and A. Steger. Synchrony and asynchrony in neural networks. Manuscript, 2009.

[6] W. Senn and R. Urbanczik. Similar nonleaky integrate-and-fire neurons with instantaneous couplings always synchronize. *SIAM J. Appl. Math.*, 61(4):1143–1155 (electronic), 2000/01.

[7] C. V. Vreeswijk, L. F. Abbott, and G. B. Ermentrout. When inhibition not excitation synchronizes neural firing. *J. Comput. Neurosc.*, 1(4):313–321, 1994.

28.4.5 Quasirandomness

One of the key successes in modern algorithmics was the development of randomized algorithms. These are algorithms that not only use deterministic computations, but that are allowed to "flip a coin" and make decisions relying on the outcome of this random experiment. While counter-intuitive to most people's thinking of computing, this simple idea became a key ingredient in many algorithms used today, mostly without the user being aware of this.

The paradigm of quasirandomness suggests to imitate a particular property of a random process deterministically. The hope is that this particular property is responsible for usefulness of the random object, and that by deterministically enforcing this property, we obtain even better results, or similar result without the risk of failure, which is inherent in many randomized constructions.

A prominent example of a highly successful application of this approach in mathematics are low-discrepancy point sets and Quasi-Monte Carlo Methods (see, e.g., Niederreiter [2]). Noting that random sample point sets work well in numerical integration, and further, that such point sets enjoy certain properties of being evenly distributed, led to the definition of low-discrepancy point sets, which proved to be superior to random sample points in numerical integration.

In this section, we present two quasirandom processes that are relevant in computer science. The rotor-router model first introduced in the physics community by Priezzhev, Dhar, Dhar, and Krishnamurthy [3] and later popularized by the mathematician James Propp [1] is a quasirandom analogue of random walks, which are the heart of many randomized algorithms. A quasirandom version of the classical "randomized rumor spreading" protocol to broadcast information in networks was developed by our group. Both processes show

interesting advantages over their random counterparts. The results now present on both models indicate that quasirandom approaches are useful in computer science as well, and that they might be a direction to continue the success story of randomized methods in algorithmics.

References

[1] A. E. Holroyd, L. Levine, K. Mészáros, Y. Peres, J. Propp, and D. B. Wilson. Chip-firing and rotor-routing on directed graphs. To appear in "In and Out of Equilibrium II," Eds. V. Sidoravicius and M. E. Vares, Birkhäuser.

[2] H. Niederreiter. *Random number generation and quasi-Monte Carlo methods.* SIAM, Philadelphia, PA, USA, 1992.

[3] V. B. Priezzhev, D. Dhar, A. Dhar, and S. Krishnamurthy. Eulerian walkers as a model of self-organized criticality. *Phys. Rev. Lett.*, 77(25):5079–5082, 1996.

The Rotor-router Model (Propp-Machine)

Investigators: Benjamin Doerr and Tobias Friedrich in cooperation with Joshua Cooper (South Carolina), Gábor Tardos (Simon Fraser University and Rényi Institute of the Hungarian Academy of Sciences), James Propp (UMass Lowell), and Joel Spencer (New York)

James Propp's rotor router model is a deterministic analogue of a random walk on a graph. Instead of distributing chips randomly, each vertex serves its neighbors in a fixed order. Cooper and Spencer [3] show a remarkable similarity of both models. If an (almost) arbitrary population of chips is placed on the vertices of a grid $\mathbb{Z}^d$ and does a simultaneous walk in the Propp model, then at all times and on each vertex, the number of chips deviates from the expected number the random walk would have gotten there, by at most a constant. This constant is independent of the starting configuration and the order in which each vertex serves its neighbors. In [2] it is shown that for the graph being the infinite path, the constant is approximately 2.3. In [4, 5] this is extended to the two-dimensional grid and it is shown that this constant is approximately 7.8 if all vertices serve their neighbors in clockwise or counterclockwise order, and 7.3 otherwise.

This result raises the question if all graphs do have this property. In [1] we are now able to answer this question negatively. For the graph being an infinite k-ary tree ($k \geq 3$), we show that for any deviation D there is an initial configuration of chips such that after running the Propp model for a certain time there is a vertex with at least D more chips than expected in the random walk model. However, to achieve a deviation of D it is necessary that at least $\exp(\Omega(D^2))$ vertices contribute by being occupied by a number of chips not divisible by k in a certain time interval.

References

- [1] J. Cooper, B. Doerr, T. Friedrich, and J. Spencer. Deterministic random walks on regular trees. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008*, San Francisco, USA, 2008, pp. 766–772. ACM.

- [2] J. Cooper, B. Doerr, J. Spencer, and G. Tardos. Deterministic random walks on the integers. *European Journal of Combinatorics*, 28(8):2072–2090, 2007.

[3] J. Cooper and J. Spencer. Simulating a random walk with constant error. *Comb. Probab. Comput.*, 15(6):815–822, 2006.

[4] B. Doerr and T. Friedrich. Deterministic random walks on the two-dimensional grid. In T. Asano, ed., *Algorithms and Computation, 17th International Symposium, ISAAC 2006*, Kolkata, India, 2006, *LNCS 4288*, pp. 474–483. Springer.

- [5] B. Doerr and T. Friedrich. Deterministic random walks on the two-dimensional grid. *Combinatorics, Probability and Computing*, 18:123–144, 2009.

Introducing Quasirandom Rumor Spreading

Investigators: Benjamin Doerr and Tobias Friedrich in cooperation with Thomas Sauerwald (International Computer Science Institute Berkeley)

The study of the dissemination of information in networks has become a topic of intense research in recent years, mainly pushed by the rapid increase network usages with the Internet just being one very visible example. A problem that often arises is that a piece of information currently held by one node of the network has to be spread to all nodes of the network as quickly and as reliably as possible.

A classical solution to this problem is the so-called *randomized rumor spreading protocol*. It proceeds in rounds. In each round, every vertex that already knows the information tells it to a neighboring node chosen at random. As a result, this node now also knows the rumor and begins to gossip in the next round.

Results of Frieze and Grimmett [3] show that this simple protocol succeeds in spreading a rumor from one node of a complete graph to all others within $O(\log n)$ rounds. For the network being a hypercube or a random graph $G(n,p)$ with $p \geq (1+\varepsilon)(\log n)/n$, also $O(\log n)$ rounds suffice [2].

In [1] we propose and analyze a quasirandom analogue of this protocol. In the quasirandom model, we assume that each node has a (cyclic) list of its neighbors. Once informed, it starts at a random position of the list, but from then on informs its neighbors in the order of the list. Surprisingly, irrespective of the orders of the lists, the above mentioned bounds still hold. In addition, we also show a $O(\log n)$ bound for sparsely connected random graphs $G(n,p)$ with $p = (\log n + f(n))/n$, where $f(n) \to \infty$ and $f(n) = O(\log \log n)$. Here, the classical model needs $\Theta(\log^2(n))$ rounds.

References

- [1] B. Doerr, T. Friedrich, and T. Sauerwald. Quasirandom rumor spreading. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008*, San Francisco, USA, 2008, pp. 773–781. ACM.

[2] U. Feige, D. Peleg, P. Raghavan, and E. Upfal. Randomized broadcast in networks. *Random Structures and Algorithms*, 1(4):447–460, 1990.

[3] A. Frieze and G. Grimmett. The shortest-path problem for graphs with random arc-lengths. *Discrete Applied Mathematics*, 10:57–77, 1985.

An Experimental Analysis of Quasirandom Rumor Spreading

Investigators: Benjamin Doerr and Tobias Friedrich in cooperation with Marvin Künnemann (Universität des Saarlandes) and Thomas Sauerwald (International Computer Science Institute Berkeley)

While the previous work shows that quasirandom rumor spreading is guaranteed to work well in spite of the greatly reduced use of randomness, the theoretical methods used there were not able to demonstrate that the quasirandom model actually gains a lot. For this reason, we experimentally compared the two models in [1].

This empirical analysis shows that the quasirandom model generally is faster (which was expected, though maybe not to this extent), but also that the runtime is more concentrated around the mean value (which is surprising given that much fewer random bits are used in the quasirandom process).

These advantages are also observed in a lossy communication model, where each transmission does not reach its target with a certain probability, and in an asynchronous model, where nodes send at random times drawn from an exponential distribution. For the theoretical background of the lossy communication model see the section about robustness below. We also show that for most graphs the particular structure of the lists has little influence on the efficiency. In particular, there is no problem if all nodes use an identical order to inform their neighbors.

References

- [1] B. Doerr, T. Friedrich, M. Künnemann, and T. Sauerwald. Quasirandom rumor spreading: An experimental analysis. In I. Finocchi and J. Hershberger, eds., *Proceedings of the 10th Workshop on Algorithm Engineering and Experiments (ALENEX 2009)*, New York, USA, 2009, pp. 145–153. SIAM.

The Evolution of Quasirandom Rumor Spreading on the Complete Graph

Investigators: Spyros Angelopoulos, Benjamin Doerr, Nikolaos Fountoulakis, Anna Huber, and Konstantinos Panagiotou in cooperation with Mahmoud Fouz and Markus Bläser (Universtität des Saarlandes)

The theoretical results described above mostly show that the quasirandom model achieves a similar runtime, where similar means identical up to constant factors. Since in a practical application also a constant-factor difference can be important, we tried to prove sharper bounds. For the classical model, only for the complete graph such precise results are known. Frieze and Grimmett [3] show that for every $\epsilon > 0$ all nodes of the complete graph on n vertices are informed within $(1 \pm \epsilon)(\log_2 n + \ln n)$ rounds with probability $1 - o(1)$.

In [1], we show that again the very same result holds for the quasirandom model, independent of the structure of the cyclic permutations. This shows that quasirandom rumor spreading is almost precisely as fast as randomized rumor spreading, *irrespective* of the lists. As a corollary we obtain that the number of random bits needed drops down from $O(\log^2 n)$ to only $O(\log n)$ at each vertex, without losing efficiency.

Whereas this reduction of random bits comes at no loss of efficiency, we further show that a subsequent reduction of randomness in a more general model will incur additional

rounds for particular choices of the lists. In this *gate model*, we assume that every vertex makes its random choice only from a subset of special vertices equidistantly distributed in its list. We prove that if ℓ is the distance between two gates, then with high probability $\ell - (1-\epsilon)(\log_2 \ell + \ln \ell)$ additional rounds are needed to inform all vertices.

Pittel [4] proved that the randomized protocol on the complete graph on n vertices informs all vertices within $\log_2 n + \ln n \pm \omega(n)$ stages with probability $1 - o(1)$, for every function $\omega(n)$ such that $\omega(n) \to \infty$ as $n \to \infty$.

In [2], we present a detailed analysis of the evolution of the quasirandom protocol on the complete graph with n vertices and show that it evolves essentially in the same way as the randomized protocol. In particular, if $S(n)$ denotes the number of stages that are needed until all vertices are informed, we show that for any slowly growing function $\omega(n)$ with probability $1 - o(1)$ we have

$$\log_2 n + \ln n - 4 \ln \ln n \leq S(n) \leq \log_2 n + \ln n + \omega(n).$$

At a high level, during each one of the first $\log_2 n$ stages, the number of informed vertices almost doubles becoming approximately equal to $n/\omega(n)$. Thereafter, there is an intermediate phase during which the number of informed vertices becomes almost equal to $n - ne^{-\omega(n)}$. Finally, there is a phase of $\ln n$ stages where the number of uninformed vertices decreases by a factor of approximately $1/e$ at each stage.

A question for further research is whether a corresponding lower bound also holds, that is, whether one can replace the $4 \ln \ln n$ term by $\omega(n)$.

References

[1] S. Angelopoulos, M. Bläser, B. Doerr, M. Fouz, A. Huber, and K. Panagiotou. Tight bounds for quasirandom rumor spreading. Submitted, 2009.

[2] N. Fountoulakis and A. Huber. Quasirandom broadcasting on the complete graph is as fast as randomized broadcasting. 2009.

[3] A. Frieze and G. Grimmett. The shortest-path problem for graphs with random arc-lengths. *Discrete Applied Mathematics*, 10:57–77, 1985.

[4] B. Pittel. On spreading a rumor. *SIAM Jour. on Appl. Math.*, 47(1):213–223, 1987.

Robustness of Quasirandom Rumor Spreading on the Complete Graph

Investigators: Benjamin Doerr, Anna Huber, and Ariel Levavi

The assumption of both the randomized and the quasirandom rumor spreading model that each transmission is surely successful is not necessarily realistic. Therefore we are currently investigating corresponding probabilistic models. These probabilistic models of the Rumor-Spreading problem are identical to the standard ones with one exception: At each attempt to contact a new node, the rumor will only be successfully transmitted with probability p. We assume that p is a constant independent from n. We prove that with probability $1-o(1)$, all the nodes in the complete graph on n vertices will be informed in $\left(\log_{1+p} n + \frac{1}{p} \ln n\right)(1 + o(1))$ steps. Together with a corresponding lower bound, this will then show that quasirandom rumor spreading is not only as fast as randomized rumor spreading, but also as robust.

28.4.6 Randomized Roundings

Randomized rounding is an important part of the basic toolbox of randomized computation, and a widely used tool in theoretical analysis. In its most basic version, the term refers to rounding a real-valued variable x to a random integer variable y such that $|x - y| < 1$ and $E[y] = x$. If this is done independently at random for many variables, then by so-called large deviation bounds, we know that with high probability, the sum of a set of variables is unlikely to change by very much through the rounding. For a typical example, assume the linear programming case of an $m \times n$ matrix A with entries in $[0, 1]$, and vectors $x \in [0, 1]^n$, $b \in \mathbb{R}^m$ such that $Ax \leq b$ holds. Then with high probability, the produced vector y satisfies

$$(Ay)_i \leq b_i + O(\sqrt{\max\{b_i, \log(m)\} \log(m)}) \tag{28.1}$$

for all $1 \leq i \leq m$. This method was derandomized by Raghavan [2], who gave a deterministic algorithm for the case that A has entries in $\{0, 1\}$ only, which in time $O(mn)$ produces a vector $y \in \{0, 1\}^n$ for which (28.1) is guaranteed to hold.

However, despite the usefulness of the method in theoretical work, virtually no experimental analysis of either randomized rounding or its derandomization has been made. One goal of the work in this section is to address this lack, and through the approach of algorithm engineering develop both the theory and the practice of randomized rounding, both in general and for particular problems. The work also addresses variants of randomized rounding where certain constraints (so-called *hard constraints*), such as the exact sum of a set of variables, hold with a guarantee, while bounds such as (28.1) continue to hold for some *soft* constraints (see e.g. [3, 1]). The presence of such hard constraints has been shown to extend the applicability of randomized rounding to more problems.

References

[1] B. Doerr. Roundings respecting hard constraints. In V. Diekert and B. Durand, eds., *STACS 2005, 22nd Annual Symposium on Theoretical Aspects of Computer Science (STACS'05)*, Stuttgart, Germany, 2005, *LNCS 3404*, pp. 617–628. Springer.

[2] P. Raghavan. Probabilistic construction of deterministic algorithms: Approximating packing integer programs. *J. Comput. Syst. Sci.*, 37:130–143, 1988.

[3] A. Srinivasan. Distributions on level-sets with applications to approximations algorithms. In *Proc. 42nd Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*, 2001, pp. 588–597.

Rounding with Constraints

Investigators: Benjamin Doerr and Magnus Wahlström

Methods for randomized rounding respecting a hard cardinality constraint have been proposed by Srinivasan [5] and Doerr [1]. Though the final result is similar in both cases—a randomized algorithm that in time $O(n)$ produces a rounding of a set of variables x such that the sum over all variables is preserved, and such that the bound (28.1) continues to hold with high probability; and in the case of Doerr's method, additionally a derandomization running in time $O(nm)$—the principles used and the resulting actual algorithms are quite different. In the work presented in [3], we initiated an experimental comparison of these methods and presented theoretical developments motivated by the results.

New Derandomizations. Derandomized versions of randomized algorithms are interesting both for theoretical reasons, and for practical reasons of solution quality (see below). Srinivasan's algorithm for producing randomized roundings guaranteeing a hard cardinality constraint [5] was presented without a derandomization, and considered by some to be unlikely to be derandomizable (due to the highly sequential nature of the random experiment). In [3] we disprove this by providing a derandomization of this method. The derandomization runs in time $O(nm)$ for n variables and m soft constraints, and guarantees the usual error bound of $\sqrt{(n/2)\log m}$. Experiments show that it performs well in practice as well, displaying the expected improvement in quality over the randomized method. The only caveat is that some care must be taken in the implementation (see method comparison below).

We also present an improvement to the derandomization of the algorithm of Doerr [2], improving the constant factor of the theoretical error bound. In experiments, this improvement is shown to reduce the rounding error by approximately a third compared to the old derandomization.

Randomized versus Derandomized Rounding. It is a well-known phenomenon among practitioners that derandomized algorithms, due to the greedy nature of the derandomized procedures, often produce better results than the randomized variants, even though the theoretical bounds are usually the same. However, this claim has never been tested experimentally. In [3], we find the claim to consistently hold for derandomized rounding with and without a hard cardinality constraint, over a range of inputs, with the advantage of the derandomization generally being a factor of 2 to 3 (bigger gaps were observed for some problems).

Also note that while creating a randomized solution takes time $O(n)$ and the derandomization procedures take time $O(nm)$ for m soft constraints, the method of generating several randomized roundings and selecting the one with the best solution still takes $O(nm)$ time per evaluation. Our results imply that if using the same amount of time to either produce one derandomized solution or select the best out of several randomly produced solutions, the quality of the derandomized solution would still be clearly better.

Method Comparison for Rounding with a Cardinality Constraint. The two approaches for generating randomized roundings respecting a cardinality constraint—that of Srinivasan [5], and that of Doerr [2]—provide no strong technical reasons to prefer the one method over the other. This holds both for the randomized approaches, and when comparing the derandomization of Srinivasan's method mentioned above to the derandomized version of Doerr's method [2]. Therefore, an experimental comparison was relevant.

We implemented randomized and derandomized versions of the two methods, as well as basic randomized rounding and Raghavan's derandomization thereof [4] for a baseline comparison, and in [3] compared the algorithms experimentally on various inputs. We refer to the article for a presentation of the results; in summary, no differences in solution quality were detected for the randomized case, while in the derandomized case, Srinivasan's algorithm generally produced somewhat better solutions than Doerr's algorithm, and also had a significantly lower running time. We have only one caveat: the derandomization of Srinivasan's algorithm was found to be sensitive to certain implementation details, where the wrong choice would lead to an algorithm that for some inputs produces consistently worse results than the randomized

version. In this regard, Doerr's method was more robust.

In general, adding a cardinality constraint did not have any negative effects on the solution quality.

References

[1] B. Doerr. Roundings respecting hard constraints. In V. Diekert and B. Durand, eds., *STACS 2005, 22nd Annual Symposium on Theoretical Aspects of Computer Science (STACS'05)*, Stuttgart, Germany, 2005, *LNCS 3404*, pp. 617–628. Springer.

[2] B. Doerr. Generating randomized roundings with cardinality constraints and derandomizations. In B. Durand and W. Thomas, eds., *STACS 2006, 23rd Annual Symposium on Theoretical Aspects of Computer Science*, Marseille, France, 2006, *LNCS 3884*, pp. 571–583. Springer.

- [3] B. Doerr and M. Wahlström. Randomized rounding in the presence of a cardinality constraint. In I. Finocchi and J. Hershberger, eds., *Proceedings of the 10th Workshop on Algorithm Engineering and Experiments (ALENEX 2009)*, New York, USA, 2009, pp. 162–174. SIAM.

[4] P. Raghavan. Probabilistic construction of deterministic algorithms: Approximating packing integer programs. *J. Comput. Syst. Sci.*, 37:130–143, 1988.

[5] A. Srinivasan. Distributions on level-sets with applications to approximations algorithms. In *Proc. 42nd Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*, 2001, pp. 588–597.

Randomized Beck-Fiala

Investigators: Benjamin Doerr, Anna Huber, and Christian Klein

The theorem of Beck and Fiala [1] yields that for every hypergraph $\mathcal{H} = (V, \mathcal{E})$ with maximum degree Δ and every set of real-valued weights $x = (x_v)_{v \in V}$ there is a rounding y of x such that $\left|\sum_{v \in E} x_v - \sum_{v \in E} y_v\right| < \Delta$ for all $E \in \mathcal{E}$. We showed that this theorem can be extended to randomized roundings, that is, we can efficiently generate the y at random so that each x_v is rounded up with probability equal to its fractional part. We can apply this result to controlled roundings and obtain good randomized roundings for matrices and higher-dimensional matrices.

References

[1] J. Beck and T. Fiala. "Integer-making" theorems. *Disc. Appl. Math.*, 3:1–8, 1981.

Construction of Uniformly Distributed Point Sets

Investigators: Benjamin Doerr and Magnus Wahlström in cooperation with Michael Gnewuch (Kiel), Peter Kritzer (University of New South Wales, Sydney), and Friedrich Pillichshammer (Johannes Kepler University Linz, Austria)

The *star discrepancy* of an n-point set T in the d-dimensional unit cube $[0,1]^d$ is given by

$$d^*_\infty(T) := \sup_{x \in [0,1]^d} \left| \frac{1}{n} |T \cap [0, x[| - vol([0, x[) \right| ,$$

where $[0, x[$ is the d-dimensional anchored half-open box $[0, x_1[\times \ldots \times [0, x_d[$. The star discrepancy is related to the worst case error of multivariate integration of a certain class of functions by quasi-Monte Carlo algorithms. Since the number of sample points is roughly proportional to the costs of these algorithms, it is of interest to find n-point configurations with small discrepancy and n not too large. In particular, the dependency of n on d should not be too bad, as there are important applications that use numerical integration in very high dimensions. For this case, bounds of the form $O(\sqrt{d/n})$, as shown (non-constructively) by Heinrich et al. [6], are more useful than more common bounds of the form $O((\log n)^d/n)$.

Construction by Rounding. In [4], the problem was attacked through a randomized rounding approach based on delta covers. Though the construction led to a bound of the form $O(\sqrt{(d \log n)/n})$, it had the disadvantages of a high running time, and a complicated algorithm for the derandomization.

Using the algorithms of [1] for generating roundings respecting hard constraints, an improved variant of the method could be presented [2], which provides the same type of discrepancy guarantee, while having an improved running time and allowing much simpler implementations.

In recent work (manuscript under preparation [5]), we perform an experimental evaluation of the quality of the point sets produced by the algorithm. Our results show that the point sets produced by our method are comparable in quality to those of more well-established, thoroughly researched methods. One complication here is the high computational difficulty of calculating or reasonably approximating the star discrepancy of a high-dimensional point set, which is a point which will be examined in future research.

Component-by-Component Point Set Construction. Because of the still relatively high time demands of the previous construction for higher numbers of dimensions, a different method, with a better time behavior, was constructed in [3]). This method works in a component-by-component fashion, adding dimensions to a point set one at a time rather than creating a point set all at once; the bound on the discrepancy of the resulting point sets is $O(\sqrt{(d^3 \log(1 + n/d))/n})$, i.e. a $d^{3/2}$-behavior rather than $d^{1/2}$, while the improvement in running time compared to [2] is essentially quadratic.

As a feature of the component-by-component approach, it is also possible to extend given high-quality low-dimensional point sets to higher dimensions by adding components to them.

An experimental evaluation of the method is under submission [5].

References

[1] B. Doerr. Generating randomized roundings with cardinality constraints and derandomizations. In B. Durand and W. Thomas, eds., *STACS 2006, 23rd Annual Symposium on Theoretical Aspects of Computer Science*, Marseille, France, 2006, *LNCS 3884*, pp. 571–583. Springer.

- [2] B. Doerr and M. Gnewuch. Construction of low-discrepancy point sets of small size by bracketing covers and dependent randomized rounding. In A. Keller, S. Heinrich, and H. Niederreiter, eds., *Monte Carlo and Quasi-Monte Carlo Methods 2006*, Ulm, Germany, 2008, pp. 299–312. Springer.

- [3] B. Doerr, M. Gnewuch, P. Kritzer, and F. Pillichshammer. Component-by-component construction of low-discrepancy point sets of small size. *Monte Carlo Methods and Applications*, 14:129–149, 2008.

[4] B. Doerr, M. Gnewuch, and A. Srivastav. Bounds and constructions for the star-discrepancy via delta-covers. *Journal of Complexity*, 21(5):691–709, 2005.

[5] B. Doerr, M. Gnewuch, and M. Wahlström. Implementation of a component-by-component algorithm to generate small low-discrepancy samples. Submitted, 2009.

[6] S. Heinrich, E. Novak, G. W. Wasilkowski, and H. Woźniakowski. The inverse of the star-discrepancy depends linearly on the dimension. *Acta Arith.*, 96:279–302, 2001.

28.5 Bio-inspired Computation

Coordinator: Frank Neumann

We consider general purpose algorithms that are often inspired by optimization processes observed in nature. Bio-inspired computation includes well-known approaches such as evolutionary algorithms (EAs) and ant colony optimization (ACO). Such algorithms can be easily applied to a wide range of problems without much problem knowledge. They are of particular interest in the case that no problem-specific algorithm is available for a given problem. There are several reason that may rule out problem-specific algorithms. On the one hand, the given problem may fall into the black box scenario, i. e., the quality of a possible solution can only be determined by running some simulation or experiment. This holds for many engineering problems. For example in automotive design the quality of a specific parameter setting can only be determined by running a simulation. In this case, evolutionary algorithms are used to design composite materials and aerodynamic shapes. On the other hand, general purpose algorithms are a good choice in practice if there are not enough resource such as time, money, or knowledge about the problem to develop good problem specific algorithms.

The design of such general purpose algorithms involves the following steps which makes them easy to apply to a given new problem.

1. Choose a representation of possible solutions.
2. Determine a function to evaluate the quality of a solution.
3. Define operators that produces from a current set of solutions a new set of solutions.

Due to the biological models that have been in mind when designing these algorithms they are not designed for the analysis in a classical sense. A major point in the design and the analysis of algorithms is to prove bounds on the runtime that such algorithms have to obtain optimal or nearly optimal solutions. Often the design of an algorithm for a certain problem is influenced by the goal of proofing that this algorithm achieves good solutions quickly. In the case of bio-inspired computation methods the goal is rather the development of algorithms that behave well for a wide range of problems by imitating optimization processes observed in nature. Such algorithms make use of many random decision especially in the construction of new solutions. They have mainly been examined experimentally to show their efficiency. Due to this background and the complex underlying stochastic processes, bio-inspired computation

methods are from a natural point of view not easy to analyze. However, there has been a lot of progress in understanding such methods rigorously in recent years. The major goal of our research is to enhance the theoretical foundations of bio-inspired computation methods by studying their computational complexity. This line of research treats bio-inspired computation methods as a class of randomized algorithms and analyzes their runtime which is measured by the number of fitness evaluations. Often the expected optimization time which refers to the expected number of fitness evaluations until an optimal solution has been reached for the first time is considered. It should be pointed out that bridging the gap between theory and practice in the field of bio-inspired computation methods is rather a long term project and will remain an important area of research for the next ten years.

Our research can be grouped into 4 subareas. We analyze basic modules of evolutionary algorithms that have a large impact on their success (Section 28.5.1) and work on evolutionary algorithms for problems from combinatorial optimization (Section 28.5.2) and multi-objective optimization (Section 28.5.3). On the other hand, we consider important aspects of ant colony optimization from a theoretical point of view (Section 28.5.4).

Cooperations and Awards

We participate in a collaborative research project on "Structural Characterization of Good Instances for Randomized Search Strategies" with IIT Kanpur and have a joint DFG project on "Theoretical Foundations of Swarm Intelligence" with DTU Copenhagen (see Section 28.16). Additionally, we cooperate with the research groups at the University of Adelaide, University of Birmingham, TU Berlin, TU Dortmund, and ETH Zurich. During the last 2 years, we won 3 best paper awards in different tracks at the Genetic and Evolutionary Computation Conference 2007 and 2008 and the overall best paper award at the International Conference on Problem Solving from Nature 2008 (see Section 28.15).

28.5.1 Design Principles of Evolutionary Algorithms

The design of an evolutionary algorithm involves several task that influence the success of these algorithms. Starting with a suitable representation of possible solutions, the designer has to think about a fitness function as well as appropriate variation operators such as crossover and/or mutation. On the other hand, selection methods play an important role as they decide which individuals constitute the individuals of the next generation. It is generally assumed that the population of an evolutionary algorithm should be a diverse set of search points of the underlying search space such that different regions of the search space are explored. Our goal is to understand the described design principles of evolutionary algorithms from a theoretical point of view by carrying out rigorous analyses. Such results give deeper insights into the impact of the mentioned modules of evolutionary algorithms.

Selection Methods

Investigators: Edda Happ, Daniel Johannsen, Christian Klein, and Frank Neumann in cooperation with Pietro S. Oliveto (University of Birmingham) and Carsten Witt (DTU

Copenhagen)

Selection methods play an important role when designing successful evolutionary algorithms. This can be observed in several applications. Rigorous runtime analyses of evolutionary algorithms mainly investigate algorithms that use elitist selection methods. Two algorithms commonly studied are Randomized Local Search (RLS) and the (1+1) EA and it is well known that both optimize any linear pseudo-Boolean function on n bits within an expected number of $O(n \log n)$ fitness evaluations [1].

We examined the use of fitness-proportional selection (also known as roulette wheel selection) which has been introduced in the context of genetic algorithms (see e.g. [3]). Here, each individual of the population is selected with a probability that depends on its fitness in correlation to the fitness of the other individuals. In [4], we analyzed variants of RLS and the (1+1) EA that use fitness proportional selection. For the analysis of RLS with fitness-proportional selection we apply results for the gambler's ruin problem [2]. To deal with the global changes imposed by the (1+1) EA, we relate the scenario of the gambler's ruin problem to a well known technique for the analysis of evolutionary algorithms called drift analysis [5]. We introduced a simplification of that method and show that fitness proportional selection leads to an exponential optimization time for any linear pseudo-Boolean function with non-zero weights with high probability. Even worse, all solutions of the algorithms during an exponential number of fitness evaluations differ with high probability in linearly many bits from the optimal solution. These theoretical studies are complemented by experimental investigations which confirm the asymptotic results on realistic input sizes.

Afterwards, we investigated in [6] how the fitness landscape influences the optimization process of population-based evolutionary algorithms using fitness-proportional selection. Considering simple pseudo-Boolean functions, we showed that it cannot be optimized in polynomial time with high probability regardless of the population size. This is proved by a generalization of drift analysis. For populations of at most logarithmic size, the negative result transfers to any function with unique optimum. Based on these insights, we investigate the effect of scaling the objective function in combination with a population that is not too small and show that then such algorithms compute optimal solutions for a wide range of problems in expected polynomial time. Finally, relationships to well-known evolutionary algorithms using elitist strategies are described.

References

[1] S. Droste, T. Jansen, and I. Wegener. On the analysis of the (1+1) evolutionary algorithm. *Theoretical Computer Science*, 276(1–2):51–81, 2002.

[2] W. Feller. *An Introduction to Probability Theory and Its Applications*, vol. 1. Wiley, 3rd edition, 1968.

[3] D. E. Goldberg. *Genetic Algorithms in Search Optimization and Machine Learning*. Addison-Wesley, 1989.

- [4] E. Happ, D. Johannsen, C. Klein, and F. Neumann. Rigorous analyses of fitness-proportional selection for optimizing linear functions. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, pp. 953–960. ACM.

[5] J. He and X. Yao. Drift analysis and average time complexity of evolutionary algorithms. *Artificial Intelligence*, 127:57–85, 2001.

• [6] F. Neumann, P. S. Oliveto, and C. Witt. Theoretical analysis of fitness-proportional selection: Landscapes and efficiency. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM. To appear.

Diversity Mechanisms

Investigators: Tobias Friedrich, Nils Hebbinghaus, and Frank Neumann in cooperation with Pietro S. Oliveto (University of Birmingham), Dirk Sudholt (TU Dortmund), and Carsten Witt (DTU Copenhagen)

It is widely assumed and observed in experiments that the use of diversity mechanisms in evolutionary algorithms may have a great impact on their running time. Maintaining diversity is important for the performance of evolutionary algorithms (see e.g. [1]). Diversity mechanisms can enhance global exploration of the search space and enable crossover to find dissimilar individuals for recombination. Up to 2007, there were no rigorous analysis pointing out how different diversity mechanisms for global exploration influence the runtime behavior. In [2], we considered evolutionary algorithms that differ from each other in the way they ensure diversity and point out situations where the right mechanism is crucial for the success of the algorithm. The considered evolutionary algorithms either diversify the population with respect to the search points or with respect to function values. Investigating simple plateau functions, we showed that using the "right" diversity strategy makes the difference between an exponential and a polynomial runtime. Later on, we examined how the drawback of the "wrong" diversity mechanism can be compensated by increasing the population size.

In [3], we further explored the global exploration capabilities of mutation-based algorithms. Using a simple bimodal test function and rigorous runtime analyses, we compared well-known diversity mechanisms like deterministic crowding, fitness sharing, and others with a plain algorithm without diversification on the simplest bimodal function that may also appear as part of a real-world problem. On the considered function, simple hill climbers find the global optimum with constant probability, hence a restart strategy is sufficient for optimization. Firstly, we rigorously proved that diversity mechanisms are necessary for our function since populations of almost linear size without diversification fail to find both peaks, with high probability. Then we analyzed the mentioned diversity mechanisms and showed that not all of them are effective for avoiding premature convergence even for such a simple landscape. As a result, we get a more objective and more general impression of the capabilities and limitations of common diversity mechanisms.

References

[1] N. Chaiyaratana, T. Piroonratana, and N. Sangkawelert. Effects of diversity control in single-objective and multi-objective genetic algorithms. *J. Heuristics*, 13(1):1–34, 2007.

• [2] T. Friedrich, N. Hebbinghaus, and F. Neumann. Comparison of simple diversity mechanisms on plateau functions. *Theoretical Computer Science*, 2009. To appear.

• [3] T. Friedrich, P. Oliveto, D. Sudholt, and C. Witt. Theoretical analysis of diversity mechanisms for global exploration. In M. Keijzer, ed., *Proceedings of the 10th annual conference on Genetic and evolutionary computation (GECCO 2008)*, Atlanta, GA, USAA, 2008, pp. 945–952. ACM.

Rang-based Mutation

Investigators: Frank Neumann in cooperation with Per Kristian Lehre (University of Birmingham) and Pietro S. Oliveto (University of Birmingham)

Determining the optimal parameters for an evolutionary algorithm is a challenging task that has been widely studied in the field of evolutionary computation [2]. There are many parameters in an evolutionary algorithm and many studies have focused on how parameters such as representation, population size or variation operator rates affect the algorithm's performance.

Recently, experimental work [1] has pointed out that it is often not useful to work with a fixed mutation rate. Therefore it was proposed that the population should be ranked according to fitness and the mutation rate of an individual should depend on its rank. The claim is that this allows the algorithm to explore new regions in the search space as well as progress quickly towards optimal solutions. Complementing these experimental investigations, we examined in [3] the proposed approach by presenting rigorous theoretical analyses. Our analyses point out the different effects that the use of rank-based mutation has on simple unimodal functions as well as on difficult deceptive trap functions. Our results for the OneMax function show that the rank-based mutation strategy is effective in climbing up slopes. Afterwards, we gave theoretical evidence of the better performance of rank-based mutation rates compared to fixed mutation rates for the trap functions considered in [1]. However, we also presented classes of functions which are deceptive for rank-based mutation leading to exponential runtime while fixed-mutation rate algorithms are efficient with high probability.

References

[1] J. Cervantes and C. R. Stephens. Rank based variation operators for genetic algorithms. In *Proc. of GECCO '08*, 2008, pp. 905–912. ACM Press.

[2] F. G. Lobo, C. F. Lima, and Z. Michalewicz, eds. *Parameter Setting in Evolutionary Algorithms*, *Studies in Computational Intelligence*, vol. 54. Springer, 2007.

- [3] P. S. Oliveto, P. K. Lehre, and F. Neumann. Theoretical analysis of rank-based mutation - combining exploration and exploitation. In *IEEE Congress on Evolutionary Computation 2009*, Trondheim, Norway, 2009. IEEE. To appear.

Neutrality

Investigators: Benjamin Doerr, Tobias Friedrich, Nils Hebbinghaus, and Frank Neumann in cooperation with Michael Gnewuch (University of Kiel)

From biology it is known that many mutations in the genotype do not have any effect on the phenotype, i.e., they are neutral. This form of redundancy was first observed by Kimura [6] when he tried to explain the high levels of polymorphism found within natural populations. The benefits of such neutral mutations have widely been discussed in the context of natural evolution (see e.g., [10, 5]). Such results from biology motivate the use of neutrality in evolutionary algorithms. Using neutrality in an evolutionary algorithm implies that additional redundancy is introduced into the considered search space. This research topic has has

attracted substantial interest in recent years. Several experimental studies have investigated whether redundancy can significantly help to come up with better algorithms [11, 9].

We presented a first rigorous runtime analysis on the effect of using neutrality in [1]. The model of neutrality analyzed in our investigations uses a single layer of constant fitness and has been introduced in [4]. In the case of the function OneMax using such a neutrality mechanism may lead to an exponential runtime while the optimization time is a small polynomial if no neutrality mechanism is used [2]. In [4] it has been stated that neutrality might be useful when considering a special class of deceptive functions. We proved that the optimization time of the (1+1) EA is exponential for all possible function values the layer can attain. Afterwards, we presented a function where neutrality is provably helpful. In particular we showed that neutrality might turn an optimization time that is exponential with probability close to 1 into an expected polynomial optimization time if the right neutrality layer is used. The analysis for our function considers the mixing time of the (1+1) EA. Using mixing time arguments similar to [7] for the analysis of evolutionary algorithms may be of independent interest and also helpful for analyzing evolutionary algorithms in other situations.

In [3], we examined the use of bit-wise neutrality which has been introduced in [8]. Bit-wise neutrality is perhaps the most simple and natural way to use neutrality when working with binary strings. In this model of neutrality the value of a phenotypic bit depends on a specific number of bits in the genome. The value of a phenotypic bit is determined by the corresponding genotypic bits and a chosen encoding function. Our investigations point out that there is a direct correlation between the mutation probability in the genotype and the phenotype for the different encoding functions investigated in [8]. Therefore working with this kind of neutrality in mutation-based evolutionary algorithms has only the effect of changing mutation probability. Due to this result it seems to be unnecessary to use bit-wise neutrality for such algorithms as the effect can also be obtained by changing the mutation probability directly in the phenotype. Later on, we pointed out that the use of bit-wise neutrality is useful when considering different numbers of genotypic bits to encode the phenotypic bits. The reason for this is that the number of genotypic bits used for a phenotypic bit determines the mutation probability for this bit in the different encodings.

References

- [1] B. Doerr, M. Gnewuch, N. Hebbinghaus, and F. Neumann. A rigorous view on neutrality. In *IEEE Congress on Evolutionary Computation 2007*, Singapore, 2007, pp. 2591–2597. IEEE.

[2] S. Droste, T. Jansen, and I. Wegener. On the analysis of the (1+1) evolutionary algorithm. *Theoretical Computer Science*, 276(1–2):51–81, 2002.

- [3] T. Friedrich and F. Neumann. When to use bit-wise neutrality. *Natural Computing*, 2009. To appear.

[4] E. Galván-López and R. Poli. Some steps towards understanding how neutrality affects evolutionary search. In *Proc. of PPSN '06*, 2006, *LNCS 4193*, pp. 778–787.

[5] M. A. Huynen. Exploring phenotype space through neutral evolution. *Journal of Molecular Evolution*, 43:165–169, 1996.

[6] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[7] I. Pak and V. H. Vu. On mixing of certain random walks, cutoff phenomenon and sharp threshold of random matroid processes. *Discrete Applied Mathematics*, 110(2-3):251–272, 2001.

[8] R. Poli and E. G. López. On the effects of bit-wise neutrality on fitness distance correlation, phenotypic mutation rates and problem hardness. In *Proc. of Foundations of Genetic Algorithms (FOGA '07)*, 2007, pp. 138–164.

[9] F. Rothlauf. Population sizing for the redundant trivial voting mapping. In *Proc. of the annual Conference on Genetic and Evolutionary Computation (GECCO '03)*, 2003, *LNCS 2724*, pp. 618–627. Springer.

[10] P. Schuster. Molecular insights into evolution of phenotypes. In J. P. Crutchfield and P. Schuster, eds., *Evolutionary Dynamics – Exploring the Interplay of Accident, Selection, Neutrality and Function*, Santa Fe Institute Series in the Science of Complexity. Oxford University Press, 2002.

[11] K. Weicker and N. Weicker. Burden and benefits of redundancy. In *Proc. of Foundations of Genetic Algorithms (FOGA '00)*, 2001, pp. 313–333. Morgan Kaufmann.

Evolutionary Algorithms and Dynamic Programming

Investigators: Benjamin Doerr and Frank Neumann in cooperation with Anton Eremeev (Sobolev Institute of Mathematics, Omsk), Christian Horoba (TU Dortmund), and Madeleine Theile (TU Berlin)

Recently, it has been shown for some combinatorial optimization problems that they can be solved by evolutionary algorithms using a suitable representation together with mutation operators adjusted to the given problem. Examples for this approach are the all-pairs shortest path problem [3] and the traveling salesman problem [4]. The representations used in these papers are different from the general encodings working with binary strings as considered earlier in theoretical works investigating the runtime behavior of evolutionary algorithms. Instead, the chosen representations reflect some properties of partial solutions of the problem at hand that allow to obtain solutions that can be extended to optimal ones for the problem at hand. To obtain such partial solutions the algorithms make use of certain diversity mechanisms that allow the algorithms proceed in a dynamic programming way [1].

Dynamic programming is a well-known algorithmic technique that helps to achieve good algorithms for a wide range of problems. A general framework for dynamic programming has been considered by Woeginger [5]. The technique allows the design of efficient algorithms, that solve the problem at hand to optimality, by extending partial solutions to optimal ones. In [2], we related the evolutionary approaches given in [3, 4] to dynamic programming and gave a general setup for evolutionary algorithms that can solve problems which have a dynamic programming formulation. In particular, we showed that any problem that can be solved by dynamic programming in time T has an evolutionary algorithm which solves the problem in expected time $O(T \cdot n \log(|DP|))$ with n being the size of the input and $|DP|$ being the size of the dynamic programming table.

References

[1] R. E. Bellman and S. E. Dreyfus. *Applied Dynamic Programming.* Princeton University Press, 1962.

- [2] B. Doerr, A. Eremeev, C. Horoba, F. Neumann, and M. Theile. Evolutionary algorithms and dynamic programming. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM. To appear.
- [3] B. Doerr, E. Happ, and C. Klein. Crossover can provably be useful in evolutionary computation. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, Proceedings of the 10th annual conference on Genetic and evolutionary computation, pp. 539–546. ACM. Best paper award.

[4] M. Theile. Exact solutions to the traveling salesperson problem by a population-based evolutionary algorithm. In *European Conference on Evolutionary Computation in Combinatorial Optimisation (EvoCOP)*, 2009. Springer. To appear.

[5] G. J. Woeginger. When does a dynamic programming formulation guarantee the existence of a fully polynomial time approximation scheme (FPTAS)? *INFORMS Journal on Computing*, 12(1):57–74, 2000.

Learning Fuzzy Rules with Evolutionary Algorithms

Investigators: Frank Neumann in cooperation with Jens Kroeske (University of Adelaide), Adam Ghandar (University of Adelaide), and Zbigniew Michalewicz (University of Adelaide)

Evolutionary algorithms combined with fuzzy rule base solution representation is a powerful real world problem solving technique, for example see [1, 6]. Fuzzy logic provides benefits in naturally representing real world quantities and relationships, fast controller adaptation, and a high capacity for solutions to be interpreted. The typical scenario involves using an EA to find optimum rule bases with respect to some application specific evaluation function, see [2, 5, 3].

The task of constructing rule base solutions includes determining rule statements, membership functions (including the number of distinct membership sets and their specific forms) and possible outputs. These parameters and the specification of data structures for computational representation have a significant impact on the characteristics and performance of the optimization process. Previous research in applications has largely consisted and relied on intuition and experimental analysis for designs and parameter settings. In [4], we took a theoretical approach to the analysis of a specific design of a fuzzy rule base optimization system that has been used in a range of successful applications; we utilized the symmetry that is inherent in the formulation to gain insight into the optimization. This leads to an interesting alternate viewpoint of the problem that may in turn lead to new approaches.

The methods we proposed can be used in an evaluation process where the error is minimized with respect to fitting rule bases to some training data. In the context of some applications, this would allow a system to learn rules with an output that is directly calculated from the data.

References

[1] A. Bagis. Determining fuzzy membership functions with tabu search - an application to control. *Fuzzy Sets and Systems*, 139(1):209–225, 2003.

[2] A. Ghandar, Z. Michalewicz, M. Schmidt, T.-D. To, and R. Zurbruegg. Computational intelligence for evolving trading rules. *IEEE Transactions on Evolutionary Computation*, 13(1):71–86, 2009.

[3] R. Johnson, M. Melich, Z. Michalewicz, and M. Schmidt. Coevolutionary optimization of fuzzy logic intelligence for strategic decision support. *IEEE Transactions on Evolutionary Computation*, 9(6):682–694, 2005.

- [4] J. Kroeske, A. Ghandar, Z. Michalewicz, and F. Neumann. Learning fuzzy rules with evolutionary algorithms - an analytic approach. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, eds., *Parallel Problem Solving from Nature (PPSN X)*, Dortmund, Germany, 2008, *LNCS 5199*, pp. 1051–1060. Springer.

[5] Z. Michalewicz, M. Schmidt, M. Michalewicz, and C. Chiriac. A decision-support system based on computational intelligence: A case study. *IEEE Intelligent Systems*, 20(4):44–49, 2005.

[6] Y. Shi, R. Eberhart, and Y. Chen. Implementation of evolutionary fuzzy systems. *IEEE Transactions on Fuzzy Systems*, 7(2):109–119, 1999.

28.5.2 Evolutionary Algorithms and Combinatorial Optimization

Evolutionary algorithms have often been shown to be successful for difficult combinatorial optimization problems appearing in various industrial, economical, and scientific domains. Our goal is to understand evolutionary algorithms in the context of combinatorial optimization problems in a rigorous way. To achieve this goal, we analyze evolutionary algorithms for problems from combinatorial optimization with respect to their runtime until they have found an optimal solution or a good approximation. These new insights should help practitioners to develop better algorithms for important difficult combinatorial optimization problems.

Shortest Path Problems

Investigators: Benjamin Doerr, Tobias Friedrich, Edda Happ, Christian Klein, and Frank Neumann in cooperation with Surender Baswana (IIT Kanpur), Somenath Biswas (IIT Kanpur), and Piyush P. Kurur (IIT Kanpur)

The first combinatorial optimization problem where rigorous runtime results have been achieved is the well-known single source shortest path (SSSP) problem [11]. Computing shortest paths in a given graph is one of the fundamental problems in computer science and still an important field of research [10, 1]. In the area of bio-inspired computation related problems such as vehicle routing [8] and routing problems in networks [7, 9] have been tackled. Therefore, it seems to be important to understand the basic SSSP problem from a theoretical point of view to gain new insights that will help practitioners solving related problems arising in applications.

In [11], the authors examined a simple EA together with a multi-objective fitness function which makes the EA mimic Dijkstra's algorithm for the SSSP problem [3]. The upper bound on the number of fitness evaluations given in that paper is $O(n^2 \ell \log n)$, where n is the number of vertices of the input graph and ℓ is the maximum over all vertices of the number of edges of a shortest path having a minimum number of edges. In [4], we improved this bound to $O(n^2 \max\{\log n, \ell\})$ and showed that this bound is tight, that is we give for each value of ℓ an example graph for which the expected optimization time is $\Omega(n^2 \max\{\log n, \ell\})$. Our bounds not only hold in expectation, but also with high probability, which is with probability $1 - O(n^{-c})$ for an arbitrary constant c. For the analysis, we used that the expected time needed to find a shortest path having ℓ' edges is $O(n^2 \ell')$. The actual time needed is

that sharply concentrated on this mean, that using a Chernoff bound and a union bound argument we get the bound of $O(n^2\ell)$. On the other hand, we get a $O(n^2 \log n)$ bound by using arguments similar to the ones used in the Coupon Collector's theorem. The lower bound can be obtained by similar arguments.

Additionally, a single-objective approach which is supposed to be efficient has been presented in [11]. However, the authors state that they were not able analyze their approach with respect to the runtime behavior. Answering the question whether the SSSP problem can be solved be a single-objective approach further insights into the optimization process of bio-inspired computation methods for this problem are needed. In [2], we showed that the single-objective approach solves the problem after $O\left(n^3 \cdot (\log n + \log w_{\max})\right)$ fitness evaluations with high probability, where $w_{\max}$ is the largest weight of the given input graph. Our proof examines the different possibilities of local changes that can decrease the distance of the current solution to an optimal one. The distance is measured as the sum of the different path lengths of the current solution minus the sum of the different paths lengths of an optimal one. We have shown that there exists at each time step a set of operations which shortens the distance by a factor of $(1 - 1/n)$. Using these insights the stated upper bounded is proven. Our analyses are complemented with a lower bound for a certain class of graphs which shows that our results are almost tight.

A generalization of the SSSP problem is the All-Pairs Shortest Path (APSP) problem. For this problem, we examined in [5] an evolutionary algorithm that uses as representation of an individual a sequence of edges. In the population we have for each pair of vertices at most one edge sequence starting in the first and ending in the second vertex. A mutation chooses one individual and adds or deletes a Poisson distributed number of times an edge at either end of the individual. We proposed three different crossover operators, which all choose two individuals at random and combine (parts of) them. If an individual is a walk, its fitness value is the length of this walk, otherwise it is infinity. A new individual is accepted, if no other individual in the population has the same start and end vertex or if the fitness of the new individual is not worse than the fitness of the other one connecting the same vertices. By arguments similar to the ones used for the SSSP problem, we showed that if only mutation is used, the optimization time of the algorithm is $\Theta(n^4)$. A rigorous analysis shows that the upper bound drops to $O(n^{3.5+\varepsilon})$ for an arbitrary $\varepsilon > 0$ if we use with any constant probability any of the crossover operators instead of the mutation operator. Later on, we improved the bound on the optimization time for the crossover-based approach to $O(n^{3.25} \log^{1/4} n)$ and showed that this bound is asymptotically tight [6]. These analyses are an important step towards understanding how crossover works and how it can be analyzed with rigorous methods. This is the first time that the usefulness of a crossover operator could be shown for a natural combinatorial problem.

References

- [1] H. Bast, S. Funke, P. Sanders, and D. Schultes. Fast routing in road networks using transit nodes. *Science*, 316(5824):566, 2007.

- [2] S. Baswana, S. Biswas, B. Doerr, T. Friedrich, P. P. Kurur, and F. Neumann. Computing single source shortest paths using single-objective fitness functions. In T. Jansen, I. Garibay,

R. Wiegand, and A. S. Wu, eds., *Proceedings of the 10th International Workshop on Foundations of Genetic Algorithms (FOGA 2009)*, Orlando, USA, 2009. ACM. To appear.

[3] E. W. Dijkstra. A note on two problems in connexion with graphs. In *Numerische Mathematik*, vol. 1, pp. 269–271. Mathematisch Centrum, Amsterdam, The Netherlands, 1959.

• [4] B. Doerr, E. Happ, and C. Klein. A tight bound for the (1+1)-ea on the single source shortest path problem. In *IEEE Congress on Evolutionary Computation 2007*, Singapore, 2007, pp. 1890–1895. IEEE.

• [5] B. Doerr, E. Happ, and C. Klein. Crossover can provably be useful in evolutionary computation. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, Proceedings of the 10th annual conference on Genetic and evolutionary computation, pp. 539–546. ACM. Best paper award.

• [6] B. Doerr and M. Theile. Improved analysis methods for crossover-based algorithms. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM. To appear.

[7] M. Dorigo and T. Stützle. *Ant Colony Optimization*. MIT Press, 2004.

[8] A. El-Fallahi, C. Prins, and R. W. Calvo. A memetic algorithm and a tabu search for the multi-compartment vehicle routing problem. *Computers & OR*, 35(5):1725–1741, 2008.

[9] S. J. Kim and M. K. Choi. Evolutionary algorithms for route selection and rate allocation in multirate multicast networks. *Appl. Intell.*, 26(3):197–215, 2007.

[10] P. Sanders and D. Schultes. Engineering highway hierarchies. In *Proc. of the 14th Annual European Symposium on Algorithms (ESA '06)*, 2006, pp. 804–816.

[11] J. Scharnow, K. Tinnefeld, and I. Wegener. The analysis of evolutionary algorithms on sorting and shortest paths problems. *Journal of Mathematical Modelling and Algorithms*, 3(4):349–366, 2004.

Sorting and Ordering

Investigators: Benjamin Doerr and Edda Happ

We introduced a new representation for sorting and ordering problems in [1]. In contrast to a previous evolutionary algorithm which uses permutations as representation [2], our representation is based on directed trees given by an array of predecessors. Given n comparable elements, we add an artificial element a_0 which is smaller than all other elements as the root of the tree. An element is supposed to be a descendant of another, if it is smaller. Thus a good initial solution is to set the predecessors of all elements to a_0. A natural mutation operator is to choose two elements having the same predecessor and to make one the new predecessor of the other. One possible fitness function is to count the number of vertex pairs for which the smaller one is a predecessor of the bigger one. If one additionally punishes vertex pairs where the opposite is the case, we assure that no such pairs are ever part of an individual. Thus, even before the algorithm has finished, a partial correct solution can be obtained.

The analysis of the $(1 + 1)$-evolutionary algorithm that arises from these parts shows that the expected optimization time is bounded by $O(n^2)$ (whereas an earlier algorithm for the sorting problem [2] has an upper bound of $O(n^2 \log n)$). Our algorithm can be efficiently implemented so that the optimization time is up to a constant factor equal to the runtime. Experiments imply that the true expected optimization time is even lower (between $O(n \log n)$

and $O(n \log^2 n)$) whereas the algorithm in [2] seems to have an expected optimization time matching the upper bound of $O(n^2 \log n)$.

References

- [1] B. Doerr and E. Happ. Directed trees: A powerful representation for sorting and ordering problems. In *Proceedings of CEC 2008*, Hong Kong, 2008, pp. 3606–3613. IEEE.

[2] J. Scharnow, K. Tinnefeld, and I. Wegener. The analysis of evolutionary algorithms on sorting and shortest paths problems. *Journal of Mathematical Modelling and Algorithms*, 3(4):349–366, 2004.

Minimum Cuts

Investigators: Frank Neumann in cooperation with Joachim Reichel (TU Berlin) and Martin Skutella (TU Berlin)

Minimum cut problems belong to the class of basic network optimization problems that occur as crucial subproblems in many real-world optimization problems and have a variety of applications in several different areas. In [3], we studied the minimum s-t-cut problem in graphs with costs on the edges in the context of evolutionary algorithms. We proved that there exist instances of the minimum s-t-cut problem that cannot be solved by standard single-objective evolutionary algorithms in reasonable time. On the other hand, we developed a bicriteria approach based on the famous MaxFlow-MinCut Theorem that enables evolutionary algorithms to find an optimum solution in expected polynomial time. The bicriteria approach takes the cost of a subset of edges as well as the remaining s-t-flow value into account that can be sent after removing the chosen edges. This trick helps to somehow enlarge the actual search space by enhancing infeasible edge sets (whose removal does not disconnect t from s). The enlarged search space no longer allows for the undesired situation in the single-objective approach discussed above.

While an explicitly given maximum s-t-flow (specified by the flow value on every edge of the graph) directly exposes a minimum s-t-cut, the maximum flow value alone does not contains any structural information about a minimum cut besides the minimum cut capacity. In particular, having access to such an oracle does not render the minimum cut problem entirely trivial. From a more practical point of view, having access to such a maximum flow oracle seems reasonable in certain situations. Consider, for example, a network of water or oil pipelines. When a leak occurs at some point t of the network, enough pipeline connections have to be cut off by using stop-cocks such that no more liquid leaks from the system. On the other hand, it is desirable to keep the number of inactivated pipeline connections at a minimum in order to keep the negative impact small. In the described scenario, after cutting off some edges, the remaining flow out of the leak can be easily observed and is actually the crucial basis for further decision-making.

Finally, in contrast to the basic minimum s-t-cut problem considered here, in more complex settings the complexity of a minimum cut computation and the related maximum flow computation can be considerably different. Consider for a example a multi-commodity flow setting with k source-sink pairs (s_i, t_i), $i = 1, \ldots, k$. Here, a maximum multi-commodity flow can be computed in polynomial time while the minimum multicut problem where it is the task to find a set of edges of minimum cost that disconnects every sink t_i from its associated source

s_i, $i = 1, \ldots, k$, is NP-hard [1]. In [2], we generalized our ideas to the NP-hard minimum multicut problem. Given a set of k terminal pairs, we proved that evolutionary algorithms in combination with a natural multi-objective model of the problem are able to obtain a k-approximation for this problem in expected polynomial time.

References

[1] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour, and M. Yannakakis. The complexity of multiterminal cuts. *SIAM J. on Comp.*, 23:864–894, 1994.

- [2] F. Neumann and J. Reichel. Approximating minimum multicuts by evolutionary multi-objective algorithms. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, eds., *Parallel Problem Solving from Nature (PPSN X)*, Dortmund, Germany, 2008, *LNCS 5199*, pp. 72–81. Springer. Best Paper Award.

- [3] F. Neumann, J. Reichel, and M. Skutella. Computing minimum cuts by randomized search heuristics. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, pp. 779–786. ACM Press.

Covering Problems

Investigators: Tobias Friedrich, Nils Hebbinghaus, Stefan Kratsch, and Frank Neumann in cooperation with Jun He (University of Wales) and Carsten Witt (DTU Copenhagen)

Covering problems occur frequently in combinatorial optimization. We mainly investigated the Vertex Cover problem. The input is given by a undirected graph $G = (V, E)$ and the task is to compute a set of vertices $V' \subseteq V$ of minimal size such that for each edge e, $e \cap V' \neq \emptyset$ holds, i.e. each edge has at least one vertex in V'. First, some simple evolutionary algorithms for single-objective optimization have been investigated. It is shown in [1] that a natural single-objective approach which minimizes the number of vertices and penalizes the number of uncovered edges has an exponential optimization even on simple bipartite graphs. One property of these bipartite graphs is that they consists of a local optimum with a large inferior neighborhood which makes it hard to obtain the global optimal solution. The other property is that these two optima differ significantly with respect to the value of their solutions. Based on these ideas it is shown that simple evolutionary algorithms can only obtain an approximation on this class of instances which is almost trivial. Based on these negative results the combination of evolutionary algorithms which classical approximation algorithms has been studied in [2]. The idea is to start with a solution which is produced by an approximation algorithm for the vertex cover problem and to improve it over time by the stochastic search process of the evolutionary algorithm. The combination of evolutionary algorithms with different approximation algorithms is investigated and the benefits and limitations of this approach are pointed out. On the other hand, we investigated in [1] how multi-objective models can enhance the optimization process for covering problems. Considering the much broader class of set covering problems, we have shown that simple evolutionary algorithms working with multi-objective models achieve a factor $\log n$-approximation in expected polynomial time.

In [3], we investigated the multi-objective model for the vertex cover problem in greater detail and related the runtime of our algorithms to the input size and the cost of an optimal solution. For the first time, it has been pointed out that the search process of evolutionary

algorithms using multi-objective models creates partial solutions that are similar to the effect of a kernelization (i. e. a special type of preprocessing from parameterized complexity). Based on this, we showed that evolutionary algorithms solve the vertex cover problem efficiently if the size of a minimum vertex cover is not too large, i.e. the expected runtime is bounded by $O(f(OPT) \cdot n^c)$, where c is a constant and f a function that only depends on OPT. This shows that evolutionary algorithms are randomized fixed-parameter tractable algorithms for the vertex cover problem.

References

- [1] T. Friedrich, J. He, N. Hebbinghaus, F. Neumann, and C. Witt. Approximating covering problems by randomized search heuristics using multi-objective models. In D. Thierens, ed., *Genetic and Evolutionary Computation Conference 2007*, London, UK, 2007, pp. 797–804. ACM.
- [2] T. Friedrich, J. He, N. Hebbinghaus, F. Neumann, and C. Witt. Analyses of simple hybrid algorithms for the vertex cover problem. *Evolutionary Computation,*, 17(1), 2009.
- [3] S. Kratsch and F. Neumann. Fixed-parameter evolutionary algorithms and the vertex cover problem. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM.

28.5.3 Evolutionary Multi-objective Optimization

Multi-objective optimization problems are often difficult to solve as the task is not to compute a single optimal solution but a set of solutions representing the different trade-offs with respect to the given objective functions. The number of these trade-offs can be exponential with regard to the problem size, which implies that not *all* trade-offs can be computed efficiently. In this case, one is interested in good approximations of the Pareto front consisting of a not too large set of Pareto-optimal solutions. It has been observed empirically that multi-objective evolutionary algorithms (MOEAs) are able to obtain good approximations for a wide range of multi-objective optimization problems. Our aim of is to contribute to the theoretical understanding of MOEAs in particular with respect to their approximation behavior.

Additional Objectives

Investigators: Tobias Friedrich, Nils Hebbinghaus, Christian Klein, and Frank Neumann in cooperation with Dimo Brockhoff (ETH Zurich) and Eckart Zitzler (ETH Zurich)

Most studies on evolutionary multi-objective optimization investigate problems where the number of considered objectives is low, i.e., between two and four, while studies with many objectives are rare, cf. [2]. The reason is that a large number of objectives leads to further difficulties with respect to decision making, visualization, and computation. Nevertheless, from a practical point of view it is desirable with most applications to include as many objectives as possible without the need to specify preferences among the different criteria. An open question in this context is how the inclusion of additional objectives affects the search efficiency of an evolutionary algorithm to generate the set of Pareto optimal solutions.

In [1], we examined how adding objectives to a given optimization problem affects the computational effort required to generate the set of Pareto optimal solutions. Experimental

studies show that additional objectives may change the running time behavior of an algorithm drastically. Often it is assumed that more objectives make a problem harder as the number of different trade-offs may increase with the problem dimension. We show that additional objectives, however, may be both beneficial and obstructive depending on the chosen objective. Our results are obtained by rigorous running time analyses that show the different effects of adding objectives to a well-known plateau-function. Additional experiments show that the theoretically shown behavior can be observed for problems with more than one objective.

In [3], we pointed out a different obstacle when using multi-objective models for single-objective optimization problems. To the best of our knowledge, there is so far no rigorous analysis of a problem on which the multi-objective approach is slower by more than a factor bounded by the population size compared to the respective single-objective one. We showed that a multi-objective model may lead to a totally inefficient optimization process (in comparison to a single-objective one) even if the population size is always small. This effect is first illustrated for simple plateau functions and later pointed out for a multi-objective model of the SetCover problem.

References

- [1] D. Brockhoff, T. Friedrich, N. Hebbinghaus, C. Klein, F. Neumann, and E. Zitzler. On the effects of adding objectives to plateau functions. *IEEE Transactions on Evolutionary Computation*, 2009. To appear.

[2] C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation)*. Kluwer Academic Publishers, New York, USA, 2002.

- [3] T. Friedrich, N. Hebbinghaus, and F. Neumann. Plateaus can be harder in multi-objective optimization. In *IEEE Congress on Evolutionary Computation 2007*, Singapore, Singapore, 2007, pp. 2622–2629. IEEE.

Fairness in Evolutionary Multi-objective Optimization

Investigators: Tobias Friedrich and Frank Neumann in cooperation with Christian Horoba (TU Dortmund)

Many multi-objective evolutionary algorithms give priority to regions in the decision or objective space that have been rarely explored. This leads to the use of fairness in evolutionary multi-objective optimization. The idea behind using fairness is that the number of descendants generated by individuals with certain properties should be balanced. In [1], we investigated the model of fairness introduced in [2]. The algorithms that are subject to our analyses count the number of descendants that have been generated by the individuals in the population. The first idea is to count the number of descendants with respect to the decision space, i. e., a separate counter is dedicated to each decision vector. The descendants are generated by individuals that have not produced many descendants in order to discover new regions of the decision space. This prevents individuals that have achieved less progress towards other non-dominated decision vectors from producing additional descendants. The other idea we examined is the usage of a counter with respect to the objective space. This implies that many decision vectors potentially depend on the same counter. We compared the runtime behavior

of these two variants and pointed out the differences of these two fairness mechanisms by rigorous analyses.

References

- [1] T. Friedrich, C. Horoba, and F. Neumann. Runtime analyses for using fairness in evolutionary multi-objective optimization. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, eds., *Parallel Problem Solving from Nature (PPSN X)*, Dortmund, Germany, 2008, *LNCS 5199*, pp. 671–680. Springer.

[2] M. Laumanns, L. Thiele, and E. Zitzler. Running time analysis of multiobjective evolutionary algorithms on pseudo-boolean functions. *IEEE Transactions on Evolutionary Computation*, 8(2):170–182, 2004.

Diversity Mechanisms for Evolutionary Multi-objective Optimization

Investigator: Frank Neumann in cooperation with Christian Horoba (TU Dortmund)

When using evolutionary algorithms for multi-objective optimization to approximate a large Pareto front, specific diversity mechanisms are applied to spread the individuals of the population over the whole Pareto front. In [1], we studied the concept of ϵ-dominance [3] and investigated its impact with respect to the runtime behavior. In the mentioned approach diversity is ensured by partitioning the objective space into boxes of appropriate size. The applied evolutionary algorithm is allowed to keep at most one individual of each box in its population. A usual scenario is to divide the objective space into boxes such that their number is logarithmic with respect to the number of objective vectors. We compared a variation of a simple evolutionary algorithm for multi-objective optimization using the concept of ϵ-dominance with the original algorithm. To point out situations where this concept leads provably to a better optimization process, we presented a class of instances with an exponential number of non-dominated feasible objective vectors. We showed that using the concept of ϵ-dominance a good approximation of the Pareto front is constructed efficiently while the approach not using this concept can not achieve this goal in expected polynomial time. Later on, we presented instances where the concept of ϵ-dominance prevents the algorithm from constructing good approximations of the Pareto front. For the efficient optimization of these instances it is essential that the population contains more than one individual per box to construct other individuals that are needed for a good approximation of the Pareto front. In contrast to this, we proved that the approach without using the diversity mechanism constructs the whole Pareto front in expected polynomial time.

Another popular approach to diversify the population of an evolutionary multi-objective algorithm is to use the density estimator. This diversity mechanism is used in a well-known evolutionary algorithm for multi-objective optimization called SPEA2 [4]. For each individual in the population the distances to all other individuals are computed. Based on these distances individuals are preferred for the next generation that do not belong to crowded regions of the objective space. In [2], we considered evolutionary algorithms for multi-objective optimization using this mechanism and examined when it is provably helpful to achieve a good approximation of the Pareto optimal set. Thereby, we also related it to the ϵ-dominance approach and pointed out in which situations one mechanism favors over the other.

References

- [1] C. Horoba and F. Neumann. Benefits and drawbacks for the use of epsilon-dominance in evolutionary multi-objective optimization. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, pp. 641–680. ACM Press.
- [2] C. Horoba and F. Neumann. Additive approximations of pareto-optimal sets by evolutionary multi-objective algorithms. In *Foundations of Genetic Algorithms 2009*, Orlando, USA, 2009. ACM. To appear.

[3] M. Laumanns, L. Thiele, K. Deb, and E. Zitzler. Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary Computation*, 10(3):263–282, 2003.

[4] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization. In *Proceedings of the EUROGEN*, 2002, pp. 95–100. CIMNE.

Computation of the Hypervolume

Investigators: Karl Bringmann and Tobias Friedrich

How to compare Pareto sets lies at the heart of research in multi-objective optimization. A measure that has been the subject of much recent study in evolutionary multi-objective optimization is the "hypervolume indicator" (HYP). It measures the volume of the dominated portion of the objective space and is of exceptional interest as it possesses the highly desirable feature of strict Pareto compliance [8]. We have shown in [3] that not only the the hypervolume indicator is **#P**-hard, but also most measures of unions of high-dimensional geometric objects. For rectangular boxes this is known as Klee's measure problem. [3] also presents an efficient FPRAS (fully polynomial-time randomized approximation scheme) for computing the volume of the unions of objects where one can (a) test whether a given point lies inside the object, (b) sample a point uniformly, and (c) calculate the volume of the object in polynomial time.

Most hypervolume indicator based optimization algorithms like SIBEA [7], SMS-EMOA [1, 5] or MO-CMA-ES [6] remove the solution with the smallest *contribution* to the dominated hypervolume from the population. This is usually iterated λ times until the size of the population no longer exceeds a fixed size μ. We show in [4] that this greedy selection scheme can perform arbitrarily bad and present the first hypervolume algorithm which calculates directly the contribution of every set of λ solutions. Given a population of size $n = \mu + \lambda$, our algorithm can calculate a set of $\lambda \geq 1$ solutions with minimal d-dimensional hypervolume contribution in time $On^{d/2} \log n + n^{\lambda}$ for $d > 2$. This improves all previously published algorithms by a factor of order $n^{\min\{\lambda, d/2\}}$ for $d > 3$.

The **#P**-hardness result of [3] for calculation of the hypervolume does not rule out that the hypervolume contribution is hard as well. In [2] it is shown that this problem is **#P**-hard to solve exactly and **NP**-hard to approximate by a factor of $2^{d^{1-\varepsilon}}$ for any $\varepsilon > 0$. It is also shown that even finding the solution with contribution at most $(1 + \varepsilon)$ times the minimal contribution of any solution is **NP**-hard. Though this dashes the hope for a provable efficient approximation algorithm, [2] also presents a very fast approximation algorithm for this problem. We prove that for arbitrarily given $\varepsilon, \delta > 0$ it calculates a solution with contribution at most $(1+\varepsilon)$ times the minimal contribution with probability at least $(1-\delta)$. The algorithm solves very large problem instances which are intractable for all previous algorithms (e.g., 10000 solutions in 100 dimensions) within a few seconds.

References

[1] N. Beume, B. Naujoks, and M. Emmerich. SMS-EMOA: Multiobjective selection based on dominated hypervolume. *European Journal of Operational Research*, 181(3):1653–1669, 2007.

- [2] K. Bringman and T. Friedrich. Approximating the least hypervolume contributor: NP-hard in general, but fast in practice. In *Proceedings of the 5th International Conference on Evolutionary Multi-Criterion Optimization (EMO 2009)*, Nantes, France, 2009. ACM.
- [3] K. Bringmann and T. Friedrich. Approximating the volume of unions and intersections of high-dimensional geometric objects. In S.-H. Hong, H. Nagamochi, and T. Fukunaga, eds., *Proceedings of the 19th International Symposium on Algorithms and Computation (ISAAC 2008)*, Gold Coast, Australia, 2008, *LNCS 5369*, pp. 436–447. Springer.
- [4] K. Bringmann and T. Friedrich. Don't be greedy when calculating hypervolume contributions. In T. Jansen, I. Garibay, W. R. Paul, and A. S. Wu, eds., *Proceedings of the 10th International Workshop on Foundations of Genetic Algorithms (FOGA 2009)*, Orlando, USA, 2009. ACM.

[5] M. Emmerich, N. Beume, and B. Naujoks. An EMO algorithm using the hypervolume measure as selection criterion. In *Proc. Third International Conference on Evolutionary Multi-Criterion Optimization (EMO '05)*, 2005, pp. 62–76.

[6] C. Igel, N. Hansen, and S. Roth. Covariance matrix adaptation for multi-objective optimization. *Evol. Comput.*, 15(1):1–28, 2007.

[7] E. Zitzler, D. Brockhoff, and L. Thiele. The hypervolume indicator revisited: On the design of Pareto-compliant indicators via weighted integration. In *Proc. Fourth International Conference on Evolutionary Multi-Criterion Optimization (EMO '07)*, 2007, *LNCS 4403*, pp. 862–876. Springer.

[8] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. da Fonseca. Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Trans. Evolutionary Computation*, 7(2):117–132, 2003.

Hypervolume-based Algorithms

Investigators: Tobias Friedrich and Frank Neumann in cooperation with Dimo Brockhoff (ETH Zurich) and Christian Horoba (TU Dortmund)

Indicator-based methods to tackle multi-objective problems have become popular recently, mainly because they allow to incorporate user preferences into the search explicitly. Multi-objective Evolutionary Algorithms (MOEAs) using the hypervolume indicator in particular showed better performance than classical MOEAs in experimental comparisons. In [1], the use of indicator-based MOEAs is investigated for the first time from a theoretical point of view. We carried out runtime analyses for an evolutionary algorithm with a $(\mu + 1)$-selection scheme based on the hypervolume indicator as it is used in most of the recently proposed MOEAs. Our analyses point out two important aspects of the search process. First, we examined how such algorithms can approach the Pareto front. Later on, we pointed out how they can achieve a good approximation for an exponentially large Pareto front.

In [2], we examined the hypervolume-based approach with respect to the achieved multiplicative approximation ratio for a given multi-objective problem and related it to a set of μ points on the Pareto front that achieves the best possible approximation ratio. For the class of linear functions and a class of convex functions, we proved that the hypervolume gives the best possible approximation ratio. In addition, we examined Pareto fronts of different shapes

by numerical calculations and showed where and when the approximation computed by the hypervolume is different to an optimal one.

References

- [1] D. Brockhoff, T. Friedrich, and F. Neumann. Analyzing hypervolume indicator based algorithms. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, eds., *Parallel Problem Solving from Nature (PPSN X)*, Dortmund, Germany, 2008, *LNCS 5199*, pp. 651–660. Springer.
- [2] T. Friedrich, C. Horoba, and F. Neumann. Multiplicative approximations and the hypervolume indicator. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM. To appear.

28.5.4 Ant Colony Optimization

Ant colony optimization (ACO) is another important bio-inspired approach to solve optimization problems. ACO algorithms are inspired by the search of an ant colony for a common source of food. It has been noticed that ants find very quickly a shortest path to a source of food. The information about which way to take to get to the food is distributed between the ants by leaving an information, called pheromone, on the way an ant has taken. As longer paths to the source take much more time than shorter paths, shorter paths are more often visited. This implies larger pheromone values on shorter paths after a small amount of time. In contrast to evolutionary algorithms where solutions are constructed from the current set of solutions, ACO algorithms obtain new solutions by random walks on a so-called construction graph. These random walks are influenced by the pheromone values of the edges of the underlying construction graph. Edges that correspond to "good" partial solutions should receive large pheromone values during the optimization process such that good solutions are obtained. Our goal is to understand the random process behind the optimization of ACO algorithms.

Update Schemes

Investigators: Benjamin Doerr, Daniel Johannsen, and Frank Neumann in cooperation with Dirk Sudholt (TU Dortmund) and Carsten Witt (DTU Copenhagen)

Regarding ACO, only convergence results [4] were known until 2006 and analyzing the runtime of ACO algorithms has been pointed out as a challenging task in [3]. Basic ant colony optimization (ACO) algorithms are governed by a single main parameter, the *evaporation factor*. Such algorithms successively generate candidate solutions to an optimization problem according to a distribution based on the solutions generated so far. The evaporation factor determines how fast the influence of past solutions decreases as new and better solutions are generated. First steps into analyzing the runtime of ACO algorithms have been made in [5], and, independently, the first runtime analyses of a simple ACO algorithm called 1-ANT were done at the same time in [8]. Subsequently this algorithm was further investigated for the optimization of some well-known pseudo-Boolean functions [2]. In [1], we simplified the view on these problems by an appropriate translation of the underlying model. This results in a profound simplification of the involved equations and a nonlinear rescaling of the evaporation factor. Investigating the rescaled evaporation factor allowed us to refine the

results given in [8]. In particular, we showed how the exponential runtime bound gradually changes to a polynomial bound inside the window of phase transition. A conclusion of the mentioned investigations is that 1-ANT is very sensitive to the choice of the evaporation factor ρ. Decreasing the value of ρ only by a small amount may lead to a phase transition and turn a polynomial runtime into an exponential one.

Many ACO algorithms used in applications use *best-so-far reinforcement* where in every iteration the current best-so-far solution is reinforced. In other words, in every iteration a pheromone update happens, using either the old or a newly generated best-so-far solution. We showed in [7] how these algorithms, variants of the MAX-MIN ant system (MMAS) (see [9]), can be analyzed for various example functions, including the class of unimodal functions and plateau functions. Thereby, we extended previous work by Gutjahr and Sebastiani [6]. The latter authors first analyzed MMAS variants on OneMax, LeadingOnes, and functions with plateaus. Their results and our contributions show that the impact of ρ is by far not as drastic as for the 1-ANT. When decreasing ρ, the algorithms become more and more similar to random search and the runtime on simple functions grows with $1/\rho$, but there is no phase transition for polynomially small ρ as for the 1-ANT. We also demonstrated how (a restricted formulation of) the fitness-level method can be adapted to the analysis of ACO algorithms. Finally, we presented lower bounds for ACO algorithms: a general lower bound for functions with unique optimum that grows with $1/\rho$ and an almost tight lower bound for LeadingOnes.

References

- [1] B. Doerr and D. Johannsen. Refined runtime analysis of a basic ant colony optimization algorithm. In *IEEE Congress on Evolutionary Computation 2007*, Singapore, 2007, pp. 501–507. IEEE.

- [2] B. Doerr, F. Neumann, D. Sudholt, and C. Witt. On the runtime analysis of the 1-ANT ACO algorithm. In D. Thierens, ed., *Genetic and Evolutionary Computation Conference 2007*, London, UK, 2007, pp. 33–40. ACM. Best paper award.

[3] M. Dorigo and C. Blum. Ant colony optimization theory: A survey. *Theoretical Computer Science*, 344:243–278, 2005.

[4] W. J. Gutjahr. ACO algorithms with guaranteed convergence to the optimal solution. *Information Processing Letters*, 82(3):145–153, 2002.

[5] W. J. Gutjahr. First steps to the runtime complexity analysis of Ant Colony Optimization. *Computers and Operations Research*, 35(9):2711–2727, 2008.

[6] W. J. Gutjahr and G. Sebastiani. Runtime analysis of ant colony optimization with best-so-far reinforcement. *Methodology and Computing in Applied Probability*, 10:409–433, 2008.

- [7] F. Neumann, D. Sudholt, and C. Witt. Analysis of different MMAS ACO algorithms on unimodal functions and plateaus. *Swarm Intelligence*, 3(1):35–68, 2009.

- [8] F. Neumann and C. Witt. Runtime analysis of a simple ant colony optimization algorithm. *Algorithmica*, 2009. To appear.

[9] T. Stützle and H. H. Hoos. MAX-MIN ant system. *Journal of Future Generation Computer Systems*, 16:889–914, 2000.

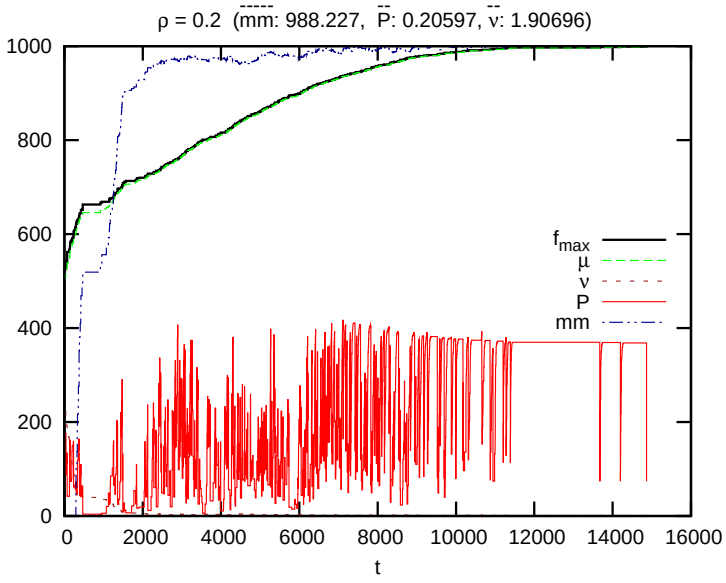


Figure 28.1: A typical run of an ACO algorithm. We track indicators of the optimization behavior like the probability to find a new solution, variance and expectation of the next solutions value, and the value of the best solution so far.

Experimental Analysis of the Optimization Behavior of Single Ant ACO Systems

Investigators: Benjamin Doerr, Daniel Johannsen, and Ching Hoo Tang

In [1], we undertook a rigorous experimental analysis of the optimization behavior of the two most studied single ant ACO systems: 1–Ant and the Max–Min Ant System (MMAS). By tracking the behavior of the underlying random processes rather than just regarding the resulting optimization time, we gained additional in sight into these systems.

These experiments are motivated and guided by preceding theoretical runtime analyses of the considered ACO systems on pseudo-boolean optimization problems. Although both, algorithms and problems, are rather basic, the rigorous theoretical analyses and especially the mathematical methods applied therein are far from simple.

To complement these theoretical results, we conducted an experimental analysis of these two single ant ACO systems on several pseudo-boolean fitness functions (onemax, leadingones, and linear functions with random weights). To gain an understanding how these algorithms work, we tracked a number of theory–guided indicators (other than the resulting optimization time) during the runs of 1–Ant and the MMAS.

It turned out that in those cases where one of the two ACO system performs well, it basically simulates the much simpler (1+1) evolutionary algorithm. This shows that the pessimistic assumptions repeatedly used in the proofs of the results mentioned above are real, and in consequence, indicates that the upper bounds on the optimization time proven there probably cannot be improved. Our analysis of the optimization behavior fits well to the fact that we rarely observed that one of the two single ant ACO systems finds the optimum

significantly faster than the (1+1) EA.

References

- [1] B. Doerr, D. Johannsen, and C. H. Tang. How single ant aco systems optimize pseudo-boolean functions. In *Parallel Problem Solving from Nature ? PPSN X*, USA, Atlanta, 2008, *LNCS 5199*, pp. 378–388. Springer.

Hybridization with Local Search

Investigators: Frank Neumann in cooperation with Dirk Sudholt (TU Dortmund) and Carsten Witt (DTU Copenhagen)

Often successful applications of ACO use a combination with local search procedures that improve the solutions constructed by the ants. The effect of using local search with ACO algorithms is manifold. Firstly, local search can help to find good solutions more quickly as it increases the "greediness" within the algorithm. Similar to memetic evolutionary algorithms, local search can also be used to discover the real "potential" of a solution as it can turn a bad looking solution into a good local optimum. Moreover, the pivot rule used in local search may guide the algorithm towards certain regions of the search space.

There is another effect that we investigated more closely in [1]. The pheromone values induce a sampling distribution over the search space. On a typical fitness landscape, once the best-so-far solution has reached a certain quality, sampling new solutions with a high variance becomes inefficient and the current best-so-far solution x^* is maintained for some time. Our analyses presented in [2] have shown that then the pheromones quickly reach the upper and lower bounds corresponding to x^*. This means that the algorithm turns to sampling close to x^*. In other words, simple ACO algorithms typically reach a situation where the "center of gravity" of the sampling distribution follows the current best-so-far solution and the variance of the sampling distribution is low.

When introducing local search into ACO algorithm, this may not be true. Local search is able to find local optima that are far away from the current best-so-far solution. In this case the "center of gravity" of the sampling distribution is far away from the best-so-far solution. We have presented simple functions where the behavior of simple ACO algorithms with and without local search have a different runtime behavior. We proved exponential runtime bounds that holds with probability exponentially close to 1 for ACO algorithms not using local search and polynomial bounds for the algorithms hybridizing ACO and local search. Our analyses exploit that the sampling distributions can follow different routes through the search space which leads to the different behavior of the two algorithmic approaches.

References

- [1] F. Neumann, D. Sudholt, and C. Witt. Rigorous analyses for the combination of ant colony optimization and local search. In M. Dorigo, M. Birattari, C. Blum, M. Clerc, T. Stützle, and A. F. T. Winfield, eds., *International Conference on Ant Colony Optimization and Swarm Intelligence 2008*, Brussels, Belgium, 2008, *LNCS 5217*, pp. 132–143. Springer.
- [2] F. Neumann, D. Sudholt, and C. Witt. Analysis of different MMAS ACO algorithms on unimodal functions and plateaus. *Swarm Intelligence*, 3(1):35–68, 2009.

28.6 Computational Geometry

Coordinators: Stefan Funke and Joachim Giesen

With the main investigators Stefan Funke and Joachim Giesen taking up chairs at other universities, the focus during the reported period for many researchers of the computational geometry group has shifted more towards geometric computing. Still, several results were obtained, mainly in the area of geometry of curves and surfaces.

28.6.1 Geometry of Curves and Surfaces

Medial Axis

Investigators: Joachim Giesen in cooperation with Balint Miklos and Mark Pauly from ETH Zürich

In [5] we had another look at the medial axis of a set of disks in the plane: given a dense sampling S of the smooth boundary of a planar shape O. We show that the medial axis of the union of Voronoi balls centered at Voronoi vertices inside O has a particularly simple structure and can be computed more efficiently and robustly than for a general union of balls.

Figure 28.2: Medial Axis of a maple leaf

Slab Support Vector Machine for Surface Reconstruction

Investigators: Joachim Giesen and Madhusudan Manjunath in cooperation with Michael Eigensatz from ETH Zürich

In [4] we analyze and compute the solution path of a parameterized optimization problem, namely the slab support vector machine (slab SVM), given a set of points in a Hilbert space that can be separated from the origin. The slab SVM aims at finding a slab (two parallel hyperplanes whose distance—the slab width—is essentially fixed) that encloses the points and is maximally separated from the origin. Extreme cases of the slab SVM include the smallest enclosing ball problem and an interpolation problem that was used (as the slab SVM itself) in surface reconstruction with radial basis functions.

On the Locality of Extracting a 2-Manifold in $\mathbb{R}^3$

Investigators: Daniel Dumitriu, Stefan Funke, and Martin Kutz in cooperation with Nikola Milosavljevic

Algorithms for reconstructing a 2-manifold from a point sample in $\mathbb{R}^3$ based on Voronoi-filtering like CRUST or CoCone still require – after identifying a set of candidate triangles – a so-called manifold extraction step which identifies a subset of the candidate triangles to form the final reconstruction surface. Non-locality of the latter step is caused by so-called slivers – configurations of four almost cocircular points having an empty circumsphere with center close to the manifold surface.

In [2, 3] we prove that under a certain mild condition – local uniformity – which typically holds in practice but can also be enforced theoretically, one can compute a reconstruction using an algorithm whose decisions about the adjacencies of a point only depend on nearby points.

While the theoretical proof requires an extremely high sampling density, our prototype implementation, described in a companion paper, performs well on typical sample sets. Due to its local mode of computation, it might be particularly suited for parallel computing or external memory scenarios.

How much Geometry it takes to Reconstruct a 2-Manifold in $\mathbb{R}^3$

Investigators: Daniel Dumitriu, Stefan Funke, and Martin Kutz in cooperation with Nikola Milosavljevic

Known algorithms for reconstructing a 2-manifold from a point sample in $\mathbb{R}^3$ are naturally based on decisions/predicates that take the geometry of the point sample into account. Facing the always present problem of round-off errors that easily compromise the exactness of those predicate decisions, an exact and robust implementation of these algorithms is far from being trivial and typically requires the employment of advanced data types for exact arithmetic as provided by libraries like CORE, LEDA or GMP. In [1] we present a new reconstruction algorithm, one of whose main novelties is to throw away geometry information early on in the reconstruction process and to mainly operate combinatorially on a graph structure. As such it is less susceptible to robustness problems due to round-off errors and also benefits from not requiring expensive exact arithmetic by faster running times. A more theoretical view on our algorithm including correctness proofs under suitable sampling conditions can be found [2].

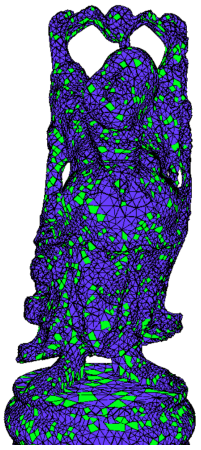


Figure 28.3: Output of our algorithm

References

- [1] D. Dumitriu, S. Funke, M. Kutz, and N. Milosavljevic. How much Geometry it takes to Reconstruct a 2-Manifold in r^3. In *10th Workshop on Algorithm Engineering and Experiments (ALENEX-2008)*, San Francisco, USA, 2008, pp. 65–74. SIAM.
- [2] D. Dumitriu, S. Funke, M. Kutz, and N. Milosavljevic. On the locality of extracting a 2-manifold in r^3. In J. Gudmundsson, ed., *11th Scandinavian Workshop on Algorithm Theory (SWAT-2008)*, Göteborg, Sweden, 2008, *LNCS 5124*, pp. 270–281. Springer.
- [3] D. Dumitriu, S. Funke, M. Kutz, and N. Milosavljevic. On the locality of extracting a 2-manifold in r^3. In S. Petitjean, ed., *Collection of abstracts of the 24th European Workshop on Computational Geometry*, Nancy, France, 2008, pp. 205–208.
- [4] M. Eigensatz, J. Giesen, and M. Manjunath. The solution path of the slab support vector machine. In P. Morin, ed., *The 20th Canadian Conference on Computational Geometry*, Mcgill University, Montreal, Canada, 2008, pp. 211–214. CCCG.

• [5] B. Miklos, J. Giesen, and M. Pauly. Medial axis approximation from inner voronoi balls: A demo of the mesecina tool. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry (SCG'07)*, Gyeongju, South Korea, 2007, pp. 123–124. ACM.

28.6.2 Misc

Epsilon-nets

Investigator: Evangelia Pyrga in cooperation with Saurabh Ray

In [2] we describe a new technique for proving the existence of small epsilon-nets for hypergraphs satisfying certain simple conditions. The technique is particularly useful for proving $o(1/\epsilon \log 1/\epsilon)$ upper bounds which the standard VC-dimension theory does not allow. We apply the technique to several geometric hypergraphs and obtain simple proofs for the existence of $O(1/\epsilon)$ size epsilon-nets for them. This includes the geometric hypergraph in which the vertex set is a set of points in the plane and the hyperedges are defined by a set of pseudo-disks.

Power Assignment Problems in Wireless Communication: Covering Points by Disks, Reaching few Receivers Quickly, and Energy-Efficient Travelling Salesman Tours

Investigators: Stefan Funke, Rouven Naujoks, and Sören Laue in cooperation with Zvi Lotker

A fundamental class of problems in wireless communication is concerned with the assignment of suitable transmission powers to wireless devices/stations such that the resulting communication graph satisfies certain desired properties and the overall energy consumed is minimized. Many concrete communication tasks in a wireless network like broadcast, multicast, point-to-point routing, creation of a communication backbone, etc. can be regarded as such a power assignment problem.

In [1] we consider several problems of that kind; the first problem was studied before and aims to select and assign powers to k out of a total of n wireless network stations such that all stations are within reach of at least one of the selected stations. We show that the problem can be $(1+\epsilon)$ approximated by only looking at a small subset of the input, which is of size $O(\frac{k^{\frac{d}{\alpha}+1}}{\epsilon^d})$, i.e. independent of n and polynomial in k and $1/\epsilon$. Here d denotes the dimension of the space where the wireless devices are distributed, so typically $d \leq 3$.

The second problem deals with the energy-efficient, bounded-hop multicast operation: Given a subset C out of a set of n stations and a designated source node s we want to assign powers to the stations such that every node in C is reached by a transmission from s within k hops. Again we show that a core set of size independent of n and polynomial in $k, |C|, 1/\epsilon$ exists, and use this to provide an algorithm which runs in time linear in n.

The last problem deals with a variant of non-metric TSP problem where the edge costs are the squared Euclidean distances; this problem is motivated by data aggregation schemes in wireless sensor networks. We show that a good TSP tour under Euclidean edge costs can be very bad in the squared distance measure and provide a simple constant approximation algorithm, partly improving upon previous results.

References

- [1] S. Funke, S. Laue, R. Naujoks, and Z. Lotker. Power assignment problems in wireless communication: Covering points by disks, reaching few receivers quickly, and energy-efficient travelling salesman tours. In S. E. Nikoletseas, B. S. Chlebus, D. B. Johnson, and B. Krishnamachari, eds., *Distributed Computing in Sensor Systems, 4th IEEE International Conference, DCOSS 2008*, Santorini Island, Greece, 2008, *LNCS 5067*, pp. 282–295. Springer.
- [2] E. Pyrga and S. Ray. New existence proofs for epsilon-nets. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, College Park, MD, USA, 2008, pp. 199–207. ACM.

28.7 Geometric Computing

Coordinator: Michael Sagraloff

The main objective is the design, analysis, and implementation of *efficient* and *reliable* geometric algorithms for complex geometric objects. Objects of interest are, for instance, *(semi-)algebraic curves, surfaces* and *arrangements* of them.

A reliable algorithm comes with some kind of guarantee on the quality of the output. Assume our goal is to compute a function $y = f(x)$. For example, for x a set of points in the plane, y could be the subset of hull vertices. Examples of guarantees are:

1. no guarantee,
2. the algorithm never crashes and always produces some output y,
3. guarantee 2 and y approximates $f(x)$ in some intuitive sense,
4. guarantee 2 and y approximates $f(x)$ is some well-defined sense,
5. guarantee 2 and $y = f(x)$.

Most existing geometric software (commercial or academic) comes with guarantee 1 or 2, ours comes with guarantee 4 or 5. What do we mean by efficiency? We want our solutions to be efficient in a double sense. First, we aim to prove that our algorithms have low complexity. Second, we want our implementations to be competitive with existing non-reliable software on inputs that can be handled by these implementations. Formulated differently, we want the running time of our algorithms to depend on the "difficulty of the input". Easy instances should be handled quickly and only hard instances should incur a higher running time.

Why is geometric computing difficult? The main reason is that geometric computing usually amounts to evaluating non-continuous functions. For example, a point lies left of, on, or right of an oriented line and hence the output of an algorithm that locates a point with respect to an oriented line belongs to a set of three discrete values $\{\textit{left of}, \textit{on}, \textit{right of}\}$. When the point and the line are given by real coordinates, we need to compute a non-continuous function on $\mathbb{R}^k$ for some k. The points of discontinuity are usually referred to as degenerate inputs. When numerical methods and floating point arithmetic are used to evaluate a geometric function, the exactness of the result cannot be guaranteed. Kettner et al. [4] have shown that even in simple cases, for example the convex hull computation of points in the plane, the presence of nearly degenerate situations, such as three nearly collinear points, may lead to

serious errors, i.e., to outputs that certainly do not satisfy guarantee 3. Also, state-of-the-art commercial CAD systems may crash when computing boolean operations of complex linear objects, see [1]. Of course, the situation becomes no simpler when dealing with non-linear objects. Computer algebra also deals with geometric objects; zero sets of polynomials are natural descriptions for curves and surfaces. The algorithms and implementations developed in computer algebra are reliable and frequently come with guarantee 5. They are however usually not efficient because of their full reliance on symbolic methods. In particular, they do not adapt to the difficulty of inputs.

In our work, we follow two paradigms for reliable geometric computing, the *Exact Geometric Computation (EGC)* paradigm and the *Controlled Perturbation (CP)* paradigm.

The *EGC approach* postulates that algorithms are complete, i.e., must handle all inputs (non-degenerate or degenerate), and that implementations ensure that all branch decisions made during the execution of an algorithm are made correctly. EGC algorithms and implementations come with guarantee 5. The term EGC was introduced by C. Yap [8]. We use the approach since the early stages of the LEDA project [5]. In linear computational geometry, EGC is a success (to which the group contributed through its work, for example, on floating-point filters, exact geometric kernels, LEDA, and CGAL). Some of the implementations have made it into products. For non-linear computational geometry, the goal is more elusive. However, we have made significant progress over the past years. We can now compute planar arrangements of algebraic curves of arbitrary degree and also arrangements induced on surfaces such as spheres and tori, see Section 28.7.3, and analyze and stratify arbitrary algebraic surfaces, see Section 28.7.4. Our success rests on the insight that approximate numerical computing can sometimes replace costly symbolic methods without sacrificing exactness. This takes the idea of floating point filters to the non-linear setting. A prime example is the Bitstream root isolator, see Section 28.7.2. On the theoretical side, we gave the first analysis of the continued fraction root isolator, see Section 28.7.2. However, despite our progress, symbolic methods are still the bottleneck in many of our algorithms, even in non-degenerate situations. For our future work, the design of symbolic/numeric algorithms with even more adaptive behavior will be crucial. We need to develop methods that solve simple cases fast and only work hard on truly hard instances. In Section 28.7.2, we report about a significant recent step in this direction. Many of our implementations have been accepted for CGAL, see Section 28.7.1.

Controlled perturbation leads to algorithms with guarantee 4. The idea is to perturb the input numerically, e.g., by adding small random noise to all input coordinates, and to solve the perturbed problem instead of the given problem. The hope is that the perturbation removes all degeneracies and puts the input into sufficiently general position so that all branch decisions can be made exactly with approximate numerical computations. The numerical computations are monitored and only relied on, if their result can be certified, e.g., by an explicit error calculation. The method [2, 3] was introduced by Dan Halperin (Tel Aviv University); he and co-workers showed that it can be used to compute arrangements of spheres and disks. We [6, 7] showed that the method is general and can be applied to a large class of geometric algorithms in the linear domain, see Section 28.7.5. Also, controlled perturbation may be a formal way of explaining, why existing geometric software using high precision floating point arithmetic works for most inputs. It is known, that some commercial systems, e.g., CPLEX, use perturbation methods.

Cooperations

During the reporting period, the Max Planck Institute for Informatics participated in the EU-funded project ACS (*Algorithms for Complex Shapes with certified numerics and topology*[1]). The project was funded for three years and ended in May 2008. It was the successor of project ECG (*Effective Computational Geometry for Curves and Surfaces*[2]). The ACS project was a cooperation with six European research groups in Groningen, Zürich (ETH), Berlin (FU), Sophia-Antipolis (INRIA), Athens, Tel-Aviv and the industrial partner GeometryFactory[3]. The aim of ACS was to advance the handling of complex shapes in geometric algorithms, both in theory and practice. For both projects about half of our resources were dedicated to software development with the focus on algebraic curves and surfaces.

Furthermore, in cooperation with the group of Dan Halperin, Tel Aviv University, we received a GIF (German-Israeli-Foundation) grant a triennial research project with the goal to develop and promote methodologies for fixed-precision approximation of complex geometric objects with quality guarantees. Its cornerstones are *Geometric Rounding, Controlled Perturbation* and *Combinatorial shape representation*. The project started in January 2009.

References

[1] P. Hachenberger and L. Kettner. Boolean operations on 3d selective Nef complexes: Optimized implementation and experiments. In L. Kobbelt and V. Shapiro, eds., *ACM Symposium on Solid and Physical Modeling (SPM 2005)*, Cambridge, MA, USA, 2005, pp. 163–174. ACM.

[2] Halperin and Shelton. A perturbation scheme for spherical arrangements with application to molecular modeling. *CGTA: Computational Geometry: Theory and Applications*, 10, 1998.

[3] D. Halperin and E. Leiserowitz. Controlled perturbation for arrangements of circles. *International Journal of Computational Geometry and Applications*, 14(4):277–310, 2004. preliminary version in SoCG 2003.

- [4] L. Kettner, K. Mehlhorn, S. Pion, S. Schirra, and C. Yap. Classroom examples of robustness problems in geometric computations. *Computational Geometry: Theory and Applications*, 40(1):61–78, 2008.

[5] K. Mehlhorn and S. Näher. The implementation of geometric algorithms. In B. Pehrson and I. Simon, eds., *Technology and foundations, Information Processing '94*, Hamburg, Germany, 1994, pp. 223–231. Elsevier.

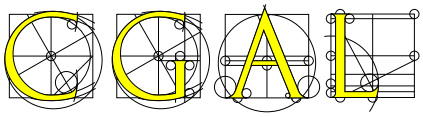
[6] K. Mehlhorn, R. Osbild, and M. Sagraloff. Reliable and efficient computational geometry via controlled perturbation. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds., *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Part I*, Venice, Italy, 2006, *LNCS 4051*, pp. 299–310. Springer.

- [7] K. Mehlhorn, R. Osbild, and M. Sagraloff. A general approach to the analysis of controlled perturbation algorithms. Technical Report ACS-TR-361502-02, University of Groningen, 9700 AB Groningen THE NETHERLANDS, 2008.

[8] C. K. Yap. Robust geometric computation. In J. E. Goodman and J. O'Rourke, eds., *Handbook of Discrete and Computational Geometry*, ch. 41, pp. 927–952. CRC Press, 2nd edition, 2004.

[1] `http://acs.cs.rug.nl/`
[2] `http://www-sop.inria.fr/prisme/ECG/`
[3] `http://www.geometryfactory.com/`

28.7.1 Software Environment



By now, most of our software is based on or is even part of CGAL, the *Computational Geometry Algorithms Library*. CGAL is the state-of-the-art in implementing geometric algorithms completely, exactly, and efficiently. However, by the time of the ECG-project, CGAL was still focused on linear objects. In particular, fundamental parts of CGAL, such as the number type support, did not suffice for dealing with algebraic curves and surfaces. Therefore, we initially concentrated our efforts into the separate software project EXACUS (*Efficient and Exact and Algorithms for Curves and Surfaces*[4]). Within the last years, EXACUS proved to be a good laboratory for testing and improving our ideas for implementations dealing with curves and surfaces. Therefore, the ACS partners appreciated our proposal to integrate core parts of EXACUS into CGAL.

Our group takes an important role within the CGAL project, which is most noticeable by the considerable number of CGAL packages that were contributed by us to the last two releases of CGAL. As a consequence of this leading role, Eric Berberich and Michael Hemmer are members of the editorial board of CGAL.

Generic Programming

In addition to the afore mentioned exact geometric computation paradigm both libraries follow the *generic programming paradigm* [1]. The paradigm suggests the gradual lifting of the concrete algorithms abstracting over details, while retaining the algorithm semantics and without compromising with its efficiency. The C++ programming language supports the generic programming paradigm by *class templates* and *function templates*. Templates are incompletely specified components where some types are only identified by formal placeholders. These types are called the *template arguments* of a template. For each instantiation of a template argument the compiler generates a separate translation of the code. Obviously, each class or function template imposes a specific set of requirements on its template arguments such as valid expressions or complexity guarantees. Usually, these requirements are grouped into so-called *Concepts*. A type that meets these requirements is called a *Model* of the concept. A concept that imposes additional requirements with respect to another concept is called a *Refinement* of that concept. For readers unfamiliar with generic programming the following analogue to mathematics should be useful: a group is a concept and ring is a refinement of this concept. A specific ring, for instance, $\mathbb{Z}$ is a model of both concepts.

Compared to object-oriented programming, the main advantage of generic programming with templates is its flexibility and efficiency. For instance, polymorphism is available without the restrictions of inheritance and does not imply a performance loss or additional memory usage due to the use of the virtual function table. Our software reflects the generic programming paradigm by the use of template techniques such as container classes, traits classes, function objects, or iterators as used in the STL [1].

[4] `http://www.mpi-sb.mpg.de/EXACUS/`

Algebraic Kernel and Curved Kernel for CGAL

Investigators: Eric Berberich, Pavel Emeliyanenko, Michael Hemmer, Michael Kerber, and Sebastian Limbach

CGAL is organized into packages and can be decomposed into three major layers. A basic layer providing fundamentals such as configuration, assertions and number types. The core part of CGAL is the geometric-kernel layer. The concept of a geometric kernel comprises all basic geometric data types and basic predicates [8, 4]. In particular, it abstracts from the choice of the used number types or filter techniques. The kernels that are provided by this layer are meant to be the main template argument to the more sophisticated data structures and algorithms in the top layer. The key idea is that the applications on the top layer use the template mechanism to instantiate their classes with the the kernel of their choice. By the beginning of the ACS project the kernels in CGAL only supported linear objects such as points, lines and triangles. As a consequence it was decided to aim for a *curved kernel* [7], which is on the same layer as the linear kernels. In addition it was decided to create an *algebraic kernel* [3] in an intermediate layer underneath. The specification splits into a univariate and a bivariate algebraic kernel, which are dedicated to solving univariate and bivariate polynomial systems, respectively. So far our efforts have led to mature implementations of the univariate kernel [6] as well as of the bivariate kernel [5]. Furthermore, a generic curved kernel has been implemented [2, 9] that relies on a model of the bivariate kernel and provides basic geometric types and predicates which can be used, for instance, to compute arrangements of curved segments.

The actual submission of our models is scheduled for next release of CGAL. A concrete application of these models can be found in Section 28.7.3.

References

[1] M. H. Austern. *Generic Programming and the STL.* Addison-Wesley, 1998.

- [2] E. Berberich and P. Emeliyanenko. CGAL's curved kernel via analysis. Technical Report ACS-TR-123203-04, Algorithms for Complex Shapes, MPI für Informatik, Germany, 2008.
- [3] E. Berberich, M. Hemmer, M. I. Karavelas, and M. Teillaud. Revision of interface specification of algebraic kernel. Technical Report ACS-TR-243301-01, University of Groningen, Groningen, The Netherlands, 2007.

[4] H. Brönnimann, A. Fabri, G.-J. Giezeman, S. Hert, M. Hoffmann, L. Kettner, S. Schirra, and S. Pion. 2d and 3d geometry kernel. In C. E. Board, ed., *CGAL User and Reference Manual*. 3.4 edition, 2008.

- [5] P. Emeliyanenko and M. Kerber. An implementation of the 2d algebraic kernel. Technical Report ACS-TR-363602-01, University of Groningen, Groningen, The Netherlands, 2008.
- [6] I. Z. Emiris, M. Hemmer, M. Karavelas, B. Mourrain, E. P. Tsigaridas, and Z. Zafeirakopoulos. Experimental evaluation and cross-benchmarking of univariate real solvers. Rapport de recherche EMIRIS:2008:INRIA-00340887:1, INRIA, Sophia Antipolis, France, 2008.

[7] I. Z. Emiris, A. Kakargias, S. Pion, M. Teillaud, and E. P. Tsigaridas. Towards an open curved kernel. In *Proc. of the 20th Annual Symposium on Comp. Geom. (SCG'04)*, 2004, pp. 438–446.

[8] A. Fabri, G. J. Giezeman, L. Kettner, S. Schirra, and S. Schönherr. On the design of CGAL, the computational geometry algorithms library. *Softw. – Pract. and Exp.*, 30(11):1167–1202, 2000.

• [9] M. Kerber. On filter methods in cgal's 2d curved kernel. Technical Report ACS-TR-243404-03, University of Groningen, Groningen, The Netherlands, 2008.

Packages Released in CGAL

Investigator: Michael Hemmer

Aiming for a support of algebraic curves and surfaces CGAL is undergoing substantial changes which in particular effects very basic layers of CGAL. During the reporting period we contributed to two releases of CGAL, release 3.3 (Jun '07) and release 3.4 (Jan '09).

- **Algebraic Foundations [1]:** This package is already part of CGAL release 3.3, it is the cornerstone for a further integration of EXACUS into CGAL. Based on the experience with the number type support in EXACUS it has enhanced the number type concepts to a more abstract level enabling the support of polynomials, finite fields, and algebraic extensions as well.
- **Number Type Support [4]:** With CGAL release 3.3, the number type support was redeveloped according to the new concept defined in the package Algebraic Foundations. This was a laborious and involved task, since it had to fulfill the new requirements while keeping backward compatibility to previous releases.
- **Modular Arithmetic [2]:** This package is part of CGAL release 3.4. The core of the package is a type representing prime fields and its smooth embedding into the remaining number type support. Modular arithmetic is a fundamental tool in modern computer algebra. In conjunction with the Chinese remainder theorem it serves as the workhorse in several algorithms such as gcd or resultant computation.
- **Polynomial [3]:** This package is part of CGAL release 3.4 and provides an implementation for dense polynomials, which is based on its EXACUS counterpart. Compared to the last EXACUS release, the class has improved in various aspects. For instance, it provides an implementation of a gcd algorithm based on modular arithmetic that allows coefficients over algebraic extension fields [5]. However, the major improvement is that the package introduces a concept for multivariate polynomials, which enables the implementation and usage of different polynomial classes.

References

• [1] M. Hemmer. *Algebraic Foundations, CGAL - Computational Geometry Algorithms Library, release 3.3.* CGAL, Campus E1 4, 66123 Saarbrücken, Germany, 2007.

• [2] M. Hemmer. *Modular Arithmetic, CGAL - Computational Geometry Algorithms Library, release 3.4.* CGAL, Campus E1 4, 66123 Saarbrücken, Germany, 2009.

• [3] M. Hemmer. *Polynomials, CGAL - Computational Geometry Algorithms Library, release 3.4.* CGAL, Campus E1 4, 66123 Saarbrücken, Germany, 2009.

• [4] M. Hemmer, S. Hert, L. Kettner, S. Pion, and S. Schirra. *Number Types, CGAL - Computational Geometry Algorithms Library, release 3.3.* CGAL, Campus E1 4, 66123 Saarbrücken, Germany, 2007.

- [5] M. Hemmer and D. Hülse. Generic implementation of a modular gcd over algebraic extension fields. In *25th European Workshop on Computational Geometry*, Brussels, Belgium, 2009, pp. 321–324. Université Libre de Bruxelles.

28.7.2 Algebraic Foundations

A key requirement in the design of ECG algorithms is the ability to perform exact computations with algebraic numbers. Such computations require a representation in purely algebraic terms, as given by the primitive element representation or Thom's encoding. A fundamental problem is *root isolation*, that is, to determine a set of disjoint, connected regions such that their union contains all roots of a univariate polynomial (or a polynomial system) and each of the regions contains precisely one root. During the report period, we made significant progress on root isolation. We proved the first polynomial bound on the complexity of the Continued Fraction method for root isolation, we derived a deterministic root isolator for polynomial with Bitstream coefficients, we developed an adaptive algorithm for determining the (complex) roots of a zero-dimensional system of integer polynomials and a more efficient method for computing subresultants of two polynomials over a domain $\mathbb{D}$. Finally, we showed how to use graphics processors for large integer arithmetic.

The Continued Fraction Algorithm

Investigator: Vikram Sharma

The continued fraction based algorithms are some of the best known algorithms for real root isolation. However, unlike other algorithms for the same task (see e.g. [6]), tight bounds on their worst case complexity are not known. In [12] we provide polynomial bounds on the worst case bit-complexity of two formulations of the continued fraction algorithm. In particular, for a square-free integer polynomial of degree n with coefficients of bit-length L, we show that the bit-complexity of Akritas' formulation is $\tilde{O}(n^8L^3)$, and the bit-complexity of a formulation by Akritas and Strzeboński is $\tilde{O}(n^7L^2)$; here $\tilde{O}$ indicates that we are omitting logarithmic factors. The analysis use a bound by Hong to compute the floor of the smallest positive root of a polynomial, which is a crucial step in the continued fraction algorithm. We also propose a modification of the latter formulation that achieves a bit-complexity of $\tilde{O}(n^5L^2)$. Our analysis sheds some light on the difficulty involved in bounding the worst case complexity of the continued fraction based algorithms.

A key component of such algorithms is a bound on the largest positive root of a polynomial. Most approaches for obtaining such bounds depend only on the absolute values of the coefficients of the polynomial. For instance, for a univariate polynomial $A(X) = \sum_{i=0}^{n} a_i X^i$, Hong's bound is

$$2 \max_{a_j<0} \quad \min_{k>j,\ a_k>0} \quad \left|\frac{a_j}{a_k}\right|^{1/(k-j)}.$$

However, as good as these bounds are, they are not known to be tight w.r.t. the largest positive root of the polynomial. Recently, we have given a general framework for obtaining bounds similar to Hong's. Within this framework, we show that Hong's bound is close to being the best bound. Moreover, all these bounds are upper bounds on the absolute positiveness of a polynomial, that is the largest positive real root amongst all the derivatives of the polynomial.

Since this, in general, may not be the positive root, we still do not know tight bounds for the positiveness of a polynomial. Finding such bounds are crucial for the continued fraction based algorithms, and may also be useful in other contexts. Even if this is not feasible, we may alternatively try to get a constant factor or a logarithmic factor approximation to the absolute positiveness of the polynomial, which would be a considerable improvement over the linear factor approximation that is obtained by Hong's bound.

Another interesting direction is to obtain a continued fraction algorithm that can handle polynomials with real numbers as coefficients, instead of integers. Or in other words, a bit-stream continued fraction algorithm. This might be achievable by modifying the current approaches for the bit-stream model to accommodate the use of the bounds mentioned above for computing subdivision points. The bit-stream approach would also lead to an improvement in both the theoretical and practical aspects of the algorithm. Thus this direction is worth investigating.

A Deterministic Bitstream Root Isolator

Investigators: Arno Eigenwillig, Kurt Mehlhorn, Michael Sagraloff, and Vikram Sharma

Root isolation is usually formulated for polynomials with integer coefficients. We also need to isolate the roots of polynomials with arbitrary real coefficients. We propose the following model. The coefficients can be approximated to any desired accuracy. In other words, the coefficients are available through their potentially infinite binary or decimal representations. However, there is no need for exact arithmetic in the field of coefficients. We coined the name Bitstream coefficients for this model.

In the previous period, we reported about a randomized subdivision algorithm for isolating the real roots of square-free polynomials with Bitstream coefficients [5]. We also reported about a partial extension to polynomials with multiple roots [4]. We continued our work on the method. In [3], a revised version of the randomized Bitstream algorithm is given with a crucial improvement in the precision management. It determines the needed precision directly, without the detour through an estimate of root separation and without the close coupling between precision and subdivision depth that existed in the former version. Furthermore, a detailed complexity analysis is given. This also includes a partial analysis of the extension to polynomials with multiple roots.

The algorithm is part of the Max Planck Institute for Informatic's model of CGAL's algebraic kernel. It is a cornerstone for our computations with curves and surfaces discussed in Sections 28.7.3 and 28.7.4. Even for polynomials with integer coefficients, it is one of the fastest methods available, see [7] for a comparison of real root solvers. The same paper recommends our solver as the method of choice for polynomials with non-rational coefficients.

We have developed a deterministic Bitstream root isolator [9, 10]. It uses the fact that the roots of a polynomial depend continuously on the coefficients in a more direct way than its randomized predecessor. The randomized algorithm worked on a polynomial with interval coefficients that represents *all* ε-approximations of the input polynomial f. In contrast, the deterministic algorithm works on a concrete ε-approximation $\tilde{f}$. Here ε is chosen such that $\tilde{f}$ and f have the same number of real roots and suitably enlarged isolating intervals of $\tilde{f}$ can serve as isolating intervals for the real roots of f. The choice of ε would be simple if the root separation of f, i.e., the smallest distance between any pair of roots, would be known. A key

part of the algorithm adapts ε as it learns about the roots of $\tilde{f}$ and hence the roots of f. The deterministic method has the same complexity as the randomized method. Moreover, the partial extension to multiple roots can be completely analyzed.

Polynomial System Solving

Investigators: Michael Hemmer, Michael Kerber, and Michael Sagraloff

Polynomial system solving is the *key* operation in non-linear computational geometry. This is due to the fact that all current algorithms in that area have shown the same behavior, namely, their bottleneck, regarding computational speed, is to find the (real) solutions of a polynomial system. Therefor we mainly distinguish three different approaches. *Exclusion or subdivision methods* continuously subdivide regions that may contain solutions, whereas regions that doubtlessly do not contain solutions are discarded. Usually, this is combined with a criterion to ensure that a region contains precisely one root, but such methods mostly fail in the presence of multiple roots (e.g. Interval Newton test) and it is hard to make them certifying. The reason is that, for the multivariate case, there exists no simple test, such as Descartes' Rule of Signs, to ensure the presence of exactly one simple root within a certain region. *Homotopy methods* [13] turned out to be a very powerful numerical approach to find the solutions of a polynomial system. They numerically track the continuous path of the known complex solutions of some trivial and appropriate polynomial system during a continuous deformation into the input system. Such methods, although very robust, lack the certification of their output in general. Finally, there exists several *elimination* methods. Multivariate resultants as well as Groebner bases are well-studied tools to obtain the solution set of a system with respect to a projection direction. In combination with a certified univariate root isolator they meet the demands in terms of the ability to certify their output. However, the coefficient explosion during their computations constitutes a severe drawback regarding the performance. A further disadvantage is that the computational costs of resultants or Groebner Basis computations depend on the bitsize and the degree of the input polynomials, but not on the real (geometric) complexity of the problem. That is, these costs are high even if the roots of the system are well separated.

Recently, we made a first step toward the design of a more adaptive algorithm [11], where we combine a subdivision-based approximation scheme for the solutions and a projection-based symbolic approach that takes place completely in several prime fields. Both approaches only deliver incomplete information, but their combination is sufficient to certify the results of the method. At the same time, the performance adaptively depends on several magnitudes in the algorithm, such as the separation of roots, instead of using worst-case bounds for them. It is planned to implement and benchmark the proposed algorithm in order to answer the question whether these advantages lead to measurable effects also in practice. The formulation of the algorithm already shows that there exist several possible optimizations of the proposed methods that an actual implementation should take care of. Some of them are easy whereas others require further studies, also within research areas not considered so far. We consider the proposed approach also suited to serve as the certification block in a hybrid approach where it is combined with a numerical method, for example, a homotopy solver to find the solutions. Finally, we would like to investigate the computation of the complexity of the proposed algorithm which we assume will also prove the adaptive behavior in theory.

Division-Free Computation of Subresultants Using Bezout Matrices

Investigator: Michael Kerber

The *subresultants* of two polynomials over a domain $\mathbb{D}$ correspond to the elements of a Euclidean remainder sequence up to a scalar factor. Since the coefficients of the subresultants never leave the ground domain $\mathbb{D}$ and their size grows moderately compared to naive pseudo-division, they have become a standard tool in Computer Algebra. The usual computation strategy for the subresultants is to iteratively perform pseudo-division, but scalar factors are divided out in each iteration to prevent a too big swell-up of the coefficient size.

A second approach to compute subresultants is by evaluating determinants of matrices. Abdeljaoued et al. [1] get the *principal subresultant coefficients* (i.e., the formal leading coefficients of the subresulants) by a modified version of the *Bezout matrix* where the principal subresultant coefficients correspond to the determinants of the $k \times k$ upper-left submatrices (also called leading principal minors). With the (sequential) *Berkowitz algorithm* [2], they compute the determinant of that matrix without divisions, and the determinants of all $k \times k$ upper-left submatrices are obtained as a by-product.

Our work [8] generalizes this approach such that the whole subresultant sequence is obtained, instead of only the principal coefficients. This is achieved by defining suitable matrices $M_1, \ldots, M_k$ that arise from the Bezout matrix by simple manipulations, so that their leading principal minors give all subresulant coefficients.

Regarding the number of arithmetic operations, the discussed method has a complexity of $O(n^5)$. This is clearly inferior to pseudo-division-based solutions whose complexity is $O(n^2)$. However, divisions in $\mathbb{D}$ might be considerably more expensive than multiplications in practice. Our experiments show that our approach is more efficient than state-of-the-art pseudo-division approaches for input polynomials with large integer coefficients, or if $\mathbb{D}$ is a polynomial ring.

References

[1] J. Abdeljaoued, G. Diaz-Toca, and L. Gonzalez-Vega. Minors of Bezout matrices, subresultants and the parameterization of the degree of the polynomial greatest common divisor. *Int. J. Comp. Math.*, 81(10):1223–1238, 2004.

[2] S. Berkowitz. On computing the determinant in small parallel time using a small number of processors. *Information Processing Letters*, 18:147–150, 1984.

- [3] A. Eigenwillig. *Real Root Isolation for Exact and Approximate Polynomials Using Descartes? Rule of Signs.* Phd thesis, Universität des Saarlandes, Saarbrücken, 2008.
- [4] A. Eigenwillig, M. Kerber, and N. Wolpert. Fast and exact geometric analysis of real algebraic plane curves. In C. W. Brown, ed., *Proceedings of the 2007 International Symposium on Symbolic and Algebraic Computation*, Waterloo, Ontario, Canada, 2007, pp. 151–158. ACM.

[5] A. Eigenwillig, L. Kettner, W. Krandick, K. Mehlhorn, S. Schmitt, and N. Wolpert. A descartes algorithm for polynomials with bit-stream coefficients. In V. G. Ganzha, E. W. Mayr, and E. V. Vorozhtsov, eds., *Computer Algebra in Scientific Computing, 8th International Workshop, CASC 2005*, Kalamata, Greece, 2005, *LNCS 3718*, pp. 138–149. Springer.

[6] A. Eigenwillig, V. Sharma, and C. K. Yap. Almost tight recursion tree bounds for the Descartes method. In J.-G. Dumas, ed., *ISSAC '06: Proceedings of the 2006 international symposium on Symbolic and algebraic computation*, Genova, Italy, 2006, pp. 71–78. ACM.

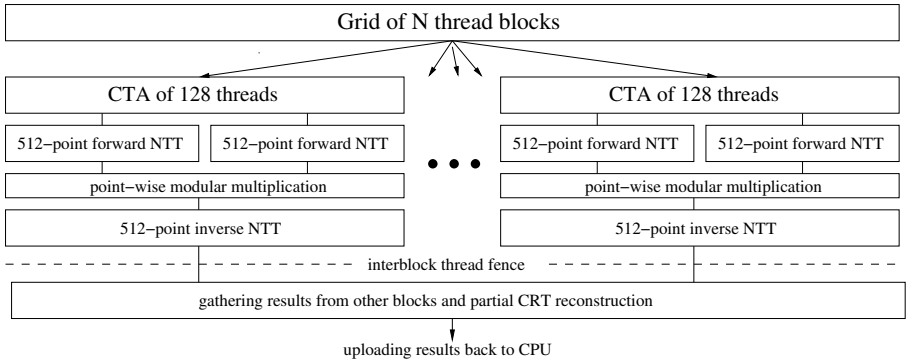


Figure 28.4: Fast modular multiplication algorithm on the GPU

- [7] I. Z. Emiris, M. Hemmer, M. Karavelas, B. Mourrain, E. P. Tsigaridas, and Z. Zafeirakopoulos. Experimental evaluation and cross-benchmarking of univariate real solvers. Rapport de recherche EMIRIS:2008:INRIA-00340887:1, INRIA, Sophia Antipolis, France, 2008.
- [8] M. Kerber. Division-free computation of subresultants using bezout matrices. International Journal of Computer Mathematics Taylor&Francis In print, 2009.
- [9] K. Mehlhorn and M. Sagraloff. A deterministic bitstream descartes algorithm. Technical Report ACS-TR-361502-03, University of Groningen, 9700 AB Groningen THE NETHERLANDS, 2008. accepted to ISSAC 2009.
- [10] K. Mehlhorn and M. Sagraloff. A deterministic descartes algorithm for real polynomials. Accepted for ISSAC 2009, 2009.

[11] M. Sagraloff, M. Kerber, and M.Hemmer. Certified complex root isolation by subdivision and modular computation. Submitted, 2009.

- [12] V. Sharma. Complexity of real root isolation using continued fractions. *Theoretical Computer Science*, 409:292–310, 2008.

[13] A. J. Sommese and C. W. Wampler. *The Numerical Solution of Systems of Polynomials Arising in Engeneering and Science*. World Scientific, Singapore, 2005.

Large Integer Arithmetic on Graphics Processors

Investigator: Pavel Emeliyanenko

From the very beginning, the graphics hardware was particularly well-suited to perform efficient computations in floating-point arithmetic. However, with the release of NVIDIA's CUDA framework [1] the situation has changed, allowing scientific applications which require integer arithmetic to benefit from the tremendous power of graphics processors.

The large integer arithmetic constitutes the core of many scientific computations. For instance, algorithms in algebraic geometry involve a large amount of symbolic computations performed over integer polynomials in one or more variables (e.g., polynomial subresultants and derived quantities). It is known that the binary segmentation method [2] reduces multiplication of polynomials with integer coefficients to one huge integer multiplication.

As it was first shown by Schönhage and Strassen [3], the Number Theoretic transform (NTT), as generalization of discrete Fourier transforms to finite fields, is asymptotically the

fastest known way to multiply two N-bit integers. They also conjectured a lower bound for multiplication as $N \log N$ which corresponds to the complexity of the Fast Fourier transform (FFT). As the NTT belongs to the family of fast orthogonal transforms, its inherent parallel structure makes it very tempting candidate for implementation on massively-threaded architectures.

The algorithm proceeds as follows: the CPU splits large integers into pieces corresponding to the size of the transform, reduces each piece modulo a set of distinct primes and loads the data on the graphics processor. The GPU launches a set of parallel NTT multiplications. In its turn, each single NTT multiplication is computed in parallel by the so-called *cooperative thread array* (CTA). Once ready, another group of thread arrays gathers multiplication results and performs partial reconstruction of the results in parallel using the *Chinese Remainder theorem* (CRT). Then, partially reconstructed results are transferred back to the CPU which computes the final product. The algorithm is schematically depicted in Figure 28.4.

Due to the limitations inferred by graphics hardware, our modular multiplication algorithm operates in a field generated by 24-bit primes which also allows for division-free modular reductions using floating-point arithmetic. The current implementation consists of highly-optimized 512-point NTT executed by a single CTA and 1024-point NTT run by two CTAs cooperatively. Our algorithm exploits redundancy in the representation of 24-bit residues with 32-bit words, in the sense that it operates on partially reduced numbers with final reductions deferred to the last stages of the algorithm.

Table to the right shows time measurements for multiplication of 16384 random integers (each of them is $43 \times 256 = 11008$ bits long) on the graphics processor using 512-point NTT with 2-steps of CRT reconstruction (which is enough to multiply integers of this bit-length) and on the CPU using GNU MP library[5].

GPU (CUDA 2.1)	39 ms
GMP 4.2.1 64-bit	837 ms
GMP 4.2.1 32-bit	1482 ms

Finally, there are several ways for future work. Efficient modular arithmetic being developed can be extended to port other algorithms which use exact arithmetic to the GPU. For instance, it enables computing polynomial subresultants by evaluating a matrix determinant modulo a set of distinct primes in parallel by graphics hardware followed by CRT reconstruction of the final result on the CPU.

References

[1] Nvidia cuda: Compute unified device architecture. NVIDIA Corp., 2007.

[2] J. von zur Gathen and J. Gerhard. *Modern Computer Algebra*. Cambridge University Press, 1999.

[3] A. Schönhage and V. Strassen. Schnelle multiplikation grosser zahlen. *Computing*, 7:281–292, 1971.

28.7.3 Arrangements of Algebraic Objects

Given a set of curves in a two-dimensional space (such as a plane) the *arrangement* induced by them is the subdivision of the space into zero-, one-, and two-dimensional cells, called vertices, edges, and faces. Reconciling exactness and efficiency in the computation of non-linear arrangements has attracted a lot of attention in recent years, which is not surprising

[5] http://gmplib.org/

due to the importance and wide applicability of arrangements in computational geometry – we just remark their importance for Constructive Solid Geometry (CSG) operations and refer to [1, 17] for further discussion. Existing inexact, floating point based implementations suffer from rounding errors which can lead to ruinous effects. In contrast, existing exact and complete methods are often not efficient enough for actual applications (e.g. [6]). Our goal is to close this gap by following an approach that is exact and complete by design, on the one hand, and efficient enough to be used in practice, on the other hand.

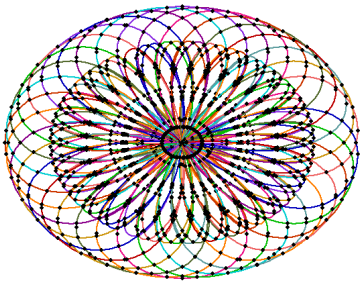
Arrangements of Algebraic Plane Curves

Investigators: Arno Eigenwillig and Michael Kerber

A popular and practical approach to compute arrangements is the sweep-line method going back to Bentley and Ottmann [3]. In [10], it is shown how to implement the basic geometric operations on points and segments needed by the sweep-line method to compute the arrangement of segments of algebraic curves of arbitrary degree (i.e., vanishing loci of polynomials $f(x, y) = \sum_{i,j} a_{ij} x^i y^j$) with integer coefficients. The approach produces the *exact* result in *all* cases, including all degeneracies, but is also fast due to a judicious combination of symbolic and adaptive-precision numeric computations. We also provide a complete implementation that will soon be released into the CGAL library (as a curved kernel; see Section 28.7.1). This is the first complete implementation of exact arrangements for such a general input type. So far, exact approaches only existed for special cases like conic curves, cubic curves, function graphs, or Bezier curves.

The sweep-line algorithm requires all input segments to be x-monotone. Therefore, all curves are initially split into x-monotone parts, using an algorithm that performs a geometric-topological analysis of the algebraic curve, called *curve analysis* [11]. The geometric predicates required by the sweep are translated into combinatorial queries on pairs of curves which can be answered by a so-called *curve pair analysis* [10]. Curve analysis and curve pair analysis are essentially equivalent to the *cylindrical algebraic decomposition* [6] of one curve or of two curves, respectively.

We tested our implementation on various benchmark instances, and compared the results with alternative approaches. As a major result, our curve analysis algorithm outperforms other algorithm for cylindrical algebraic decomposition and topology computation in the presence of singular points. The reason is that, unlike previous approaches, our algorithm employs fast numerical subroutines in each instance (without sacrificing exactness of the overall result); a fall-back to a purely symbolical method is never necessary. This behavior carries over for the curve pair analysis algorithm. The arrangement algorithm has shown to be competitive to CUBIX, a specialized arrangement algorithm for algebraic curves of degree up to three [12].



An arrangement of various rotations of a base curve of degree 4.

The implementation that we provide for exact arrangements of planar algebraic curves is generic in the coefficient type of the polynomials. Currently, it works both for integer coefficients and for square root extensions; other coefficient types can be supported by

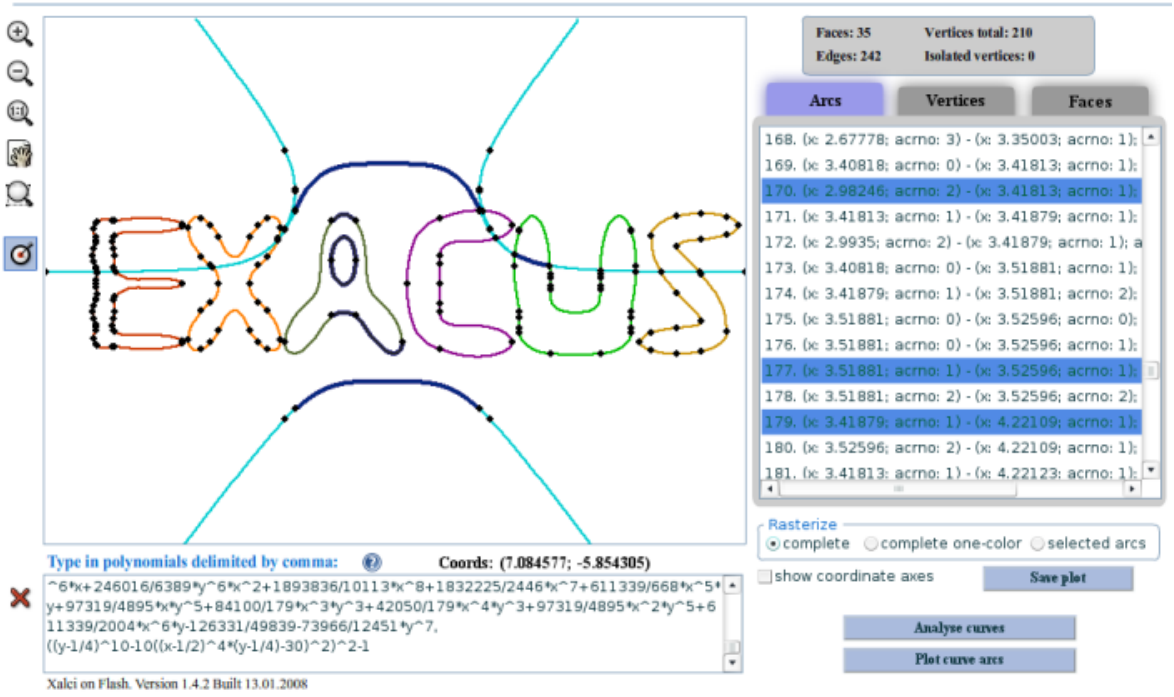


Figure 28.5: Arrangement of a set of curves with some arcs highlighted

specialization of the relevant traits classes in CGAL. Using square root extensions allows to compute arrangements of rotated curves, if the rotation angle is constructible by compass and straight-edge. Our software is based on CGAL's `Arrangement_2` package. so one has access to the full machinery of arrangements, such as overlay computation or point location, also for algebraic segments. Moreover, more sophisticated operations, such as boolean operations on polygons defined by algebraic segments, come into reach.

A worst-case complexity analysis of the arrangement algorithm is currently under consideration. For the analysis of one curve, preliminary results indicate the same upper bound as the best known complexity bounds for topology computation [7].

Remote Computation and Visualization of Algebraic Arrangements

Investigators: Pavel Emeliyanenko and Michael Kerber

To evaluate the efficiency and robustness of our software we developed a web-based application available in the Internet via `http://exacus.mpi-inf.mpg.de`. The application consists of a Macromedia Flash client, CGI-script running on Apache web-server and a multi-threaded stand-alone server carrying out all geometric computations. The video presentation about this work appears in [14].

Geometric part of the server software rests upon two algorithms implemented within the CGAL project: computing exact arrangements induced by real algebraic plane curves (Section 28.7.3) and topologically-correct rasterization of algebraic curves and their arrange-

ments [13]. It is worth mentioning that all degenerate cases (e.g., isolated vertices, vertical asymptotes, tangential intersections) are handled exactly. A combinatorial description of an arrangement along with its accurate plot is computed on the server and sent back to the Macromedia Flash[6] client to display the results in the web-browser. With this structure, the user is not burdened with installing additional software to use our application.

Once the curve arrangement is computed, its representation (combinatorial structure) is stored and cached on the server for future use. In this way, the client not only obtains a static plot of an arrangement but has a complete topologically correct geometric object at his disposal. The user can interactively explore an arrangement by zooming, panning or highlighting its features. In particular, besides an arrangement graph, the web-interface provides description of 0D, 1D and 2D arrangement components grouped in separate lists. Clicking on any item from these lists emphasizes the respective feature. In its turn, clicking an arrangement plot, executes a *point location query* on the server causing a corresponding 2D arrangement component (a face) to be highlighted, see Figure 28.5.

Accurate visualization of 1D arrangement components (algebraic curve arcs) is done on separate basis. The algorithm given in [13] is a mixture of space subdivision and curve tracking methods benefiting from advantages of both of them. To rasterize a curve arc, the algorithm chooses a "seed point" lying on this arc using exact topological description of an algebraic curve and tracks the arc in two opposite directions. In each step it encounters 8 different directions (pixels) to follow. Range analysis [21] and some heuristic observations enable to compute quickly the next pixel in a curve trace. In case of a tie, the current pixel is subdivided recursively into 4 even parts until the tracing direction can be determined.

Local space subdivision stops as soon as a certain threshold is reached (which is chosen empirically within the size of a single pixel) and all curve branches appear to leave the current pixel in a one unique direction. One might conceive of a "bunch" of curve arcs traced together until one of them goes apart. When this happens, the *real root isolation* (as given in section 28.7.2) takes place and a new seed point is selected. This technique allows to trace the majority of algebraic curves with double-precision arithmetic only, even those for which the curve branch separation is beyond the reach of floating-point accuracy. To handle highly-degenerate cases, we employ multi-precision and rational arithmetic, the most appropriate one is chosen automatically on demand.

In addition, our web-server is used to visualize and animate algebraic curves for Kempe linkage algorithm `http://www.a-kobel.de/kempe`.

Arrangement on Surfaces

Investigators: Eric Berberich, Michael Kerber, and Kurt Mehlhorn

In close collaboration with Dan Halperin's group at Tel-Aviv University, Israel, we have extended CGAL's mature `Arrangement_2` package to a framework for the construction, maintenance, and manipulation of arrangements of curves embedded on two-dimensional orientable parametric surfaces three-dimensional space. The arrangements are induced by a set of curves embedded on such a surface. Our framework applies to planes, cylinders, spheres, tori, and surfaces homeomorphic to them.

[6] http://www.macromedia.com/software/flash/about/

Part of the package are well-known paradigms to process a single curve or a set of curves: *zone and sweep line traversal*. Both were originally designed for arrangements of bounded curves in the plane. We generalized them to deal with curves embedded on parametric surfaces. To do so, we defined a compact interface for CGAL's new `Arrangement_on_surface_2` package. The new package containing our implementation of the new extensions replaces the old one. Following CGAL, it adheres to the *generic-programming* paradigm, making extensive use of class- and function-templates; see also Section 28.7.1. The framework has a compact interface: Only the operations in the interface need to be implemented for a specific application. In this way, the package can be conveniently adapted to different scenarios.

A simple scenario well supported are Voronoi diagrams in the plane. They strongly require to deal with unbounded curves. In previous approaches curves had to be clipped by the user in a pre-processing phase, so that no essential information about the arrangements (e.g., a finite intersection point) was lost. Together with a post-processing allows one to recover the unbounded faces. However doing so outside the package has the consequence that many nice functions of the package, for instance, point location or overlays, are no longer available. In contrast, our solution provides a clean and convenient access to several unbounded faces. This also enables to represent a (two-dimensional) *minimization diagram* of a set of surfaces representing their *lower envelope*: Each face is labeled with the lowest surfaces above it [22].

On surfaces, in contrast to the standard sweep-line, we have to deal with certain additional issues, that occur, as we allow that the parameterization is not injective on the boundary of the possible rectangular parameter space. That is, we first have to ensure a unique order of events in parameter space which is provided for the surface and curves in focus by a well-designed and surface-specific geometric interface. This well-known two-dimensional ground implies a parallel sweep on the surface itself: While sweep-line events in parameter-space might be unrelated, on the surface they actually refer to the same point. Our new topological interface takes care of those coincidences. It also ensures a proper nesting of components of the induced arrangement graph: If the surface is homeomorphic to a disc, the nesting is similar to the plane. On a sphere-like, cylinder-like, or torus-like surface, we maintain other natural nestings. More details appeared in [4].

There also exists concrete instantiations for the framework. For backward compatibility, the package provides the topology functions for bounded curves in the plane. But, the new default topological interface for curves in the plane allows them to be unbounded. We introduced a fictitious bounding rectangle to close *unbounded* faces. Various families of planar input curves are supported: Linear objects, circles, conics, Bézier curves, and rational function graphs. Michael Kerber et. al. provide a prototypical implementation for algebraic curves of arbitrary degree; see also Section 28.7.3. For non-planar surfaces, Tel-Aviv University maintains arrangements of geodesic arcs on a unit sphere using rational arithmetic only. In addition, we implemented proper classes to compute and query arrangements induced by algebraic surfaces intersecting (elliptic) *quadrics* (such as ellipsoids, cylinders, and paraboloids) and *ring Dupin cyclides* (a generalization of tori). The latter are surfaces of genus one. In both cases, we reduced the required geometric operations to algebraic curves in the plane, either by a projection or a direct rational parameterization. Following, their efficiency depends on the planar counterparts, while the surface-specific topological operations are not significant for the total running times. Figure 28.6 gives an arrangement on a cyclide induced by 5 algebraic surfaces of degree 3, intersecting the surface. Detailed information are published in [5].

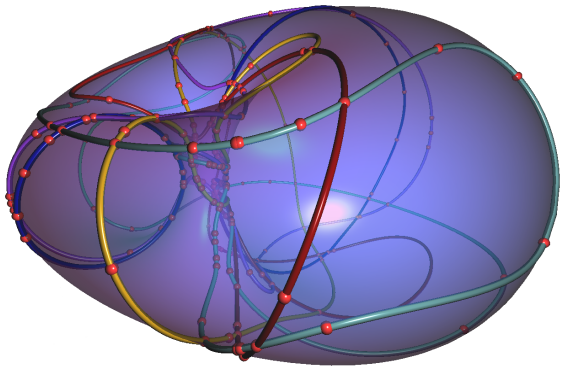


Figure 28.6: Arrangement on a cyclide. It consists of 240 vertices, 314 edges, and 74 faces. Rendering is done using the extended planar visualization algorithm [13] as previously presented.

Towards the Arrangement of Quadrics

Investigators: Michael Hemmer and Sebastian Limbach

We aim for the computation of the three-dimensional arrangement $\mathcal{A}(Q)$ induced by a given set $\mathcal{Q}$ of quadric surfaces, or quadrics for short. A *quadric* Q_f is given by a trivariate polynomial $f \in \mathbb{Q}[x, y, z]$ of degree 2. Quadrics cover a couple of common surfaces such as spheres, ellipsoids, cones, cylinders, hyperboloids, planes, and double planes.

In Hemmer et al. [8, 18], we recently reached a major goal towards the arrangement of quadrics, namely an implementation for the computation of the adjacency graph $\mathcal{G}(Q)$. The approach is based on Dupont et al. [9, 20], which provides an exact parameterization of the appearing intersection curves. Using these parameterizations, the approach represents the intersection points of quadrics by the exact parameter values with respect to the intersection curves they lie on. Therefore, it is possible to sort the points along the curve, which enables the initialization of the adjacency graph.

However, the adjacency graph $\mathcal{G}(Q)$ does not contain enough information to represent the full arrangement, since it just stores all vertices and edges of the arrangement $\mathcal{A}(Q)$. The missing parts are the two and three dimensional cells, that is, the faces and volumes of the arrangement and their incidences at each vertex. At each vertex these incidences can be represented by a two dimensional arrangement which is conceptually embedded on a sphere surrounding the vertex. We call this arrangement the *environment map* of a vertex. This is similar to the existing approach for the arrangement of planes in [16].

Following the generic programming paradigm [2] we do not enforce that such an environment map is constructed via the intersection with a certain bounding object. So far, we have investigated two approaches. The first approach computes the arrangement on a concrete sphere using the `Arrangement_on_surfaces_2`. The second approach considers the arrangement which is induced by the sufficiently good approximation of the tangent planes of the quadrics at the given vertex. The arrangement is represented using the CGAL class `Nef_polyhedron_S2` of the corresponding CGAL package [15]. Obviously, this approach is not applicable in all cases but it turned out to be a very efficient filter [19].

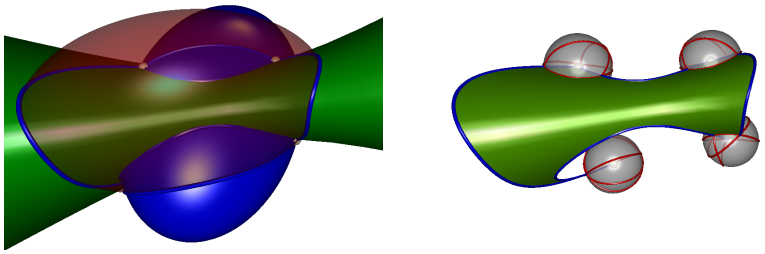


Figure 28.7: To the left: The intersection curves and vertices of two ellipsoids (blue, transparent red) and a hyperboloid of one sheet (green). To the right: The sphere maps of the four vertices and a face on the hyperboloid.

Given the information of the adjacency graph and the environment maps, it is now possible to initialize the boundary cycles of all faces of the arrangement $\mathcal{A}(Q)$. To initialize a cycle we start with an edge in an environment map that was not visited so far and follow the corresponding edge of $\mathcal{G}(Q)$ to the next vertex. At this vertex we determine the next edge of the boundary cycles within its environment map. We repeat this until we find the initial edge. A face with one boundary cycle is illustrated in Figure 28.7. So far, it is not possible to determine the nesting of boundary cycles, which is necessary to handle faces with holes, that is, faces with more than one boundary cycle. We propose to determine the nesting using "ray"-shooting on the surface of the corresponding quadric. The approach is implemented within EXACUS.

References

[1] P. K. Agarwal and M. Sharir. Arrangements and their applications. In J.-R. Sack and J. Urrutia, eds., *Handbook of Computational Geometry*, pp. 49–119. Elsevier, 2000.

[2] M. H. Austern. *Generic Programming and the STL*. Addison-Wesley, 1998.

[3] J. L. Bentley and T. A. Ottmann. Algorithms for reporting and counting geometric intersections. *IEEE Transactions on Computers*, C-28:643–647, 1979.

- [4] E. Berberich, E. Fogel, D. Halperin, K. Mehlhorn, and R. Wein. Sweeping and maintaining two-dimensional arrangements on surfaces: A first step. In L. Arge, M. Hoffmann, and E. Welzl, eds., *Algorithms - ESA 2007, 15th Annual European Symposium*, Eilat, Israel, 2007, *LNCS 4698*, pp. 645–656. Springer.

- [5] E. Berberich and M. Kerber. Exact arrangements on tori and dupin cyclides. In E. Haines and M. McGuire, eds., *Proceedings of the 2008 ACM Symposium on Solid and Physical Modeling*, Stony Brook, USA, 2008, pp. 59–66. ACM.

[6] B. F. Caviness and J. R. Johnson, eds. *Quantifier Elimination and Cylindrical Algebraic Decomposition, Texts and Monographs in Symbolic Computation*. Springer, 1998.

[7] D. I. Diochnos, I. Z. Emiris, and E. P. Tsigaridas. On the complexity of real solving bivariate systems. In *ISSAC '07: Proceedings of the 2007 international symposium on Symbolic and algebraic computation*, 2007, pp. 127–134. ACM.

- [8] L. Dupont, M. Hemmer, S. Petitjean, and E. Schömer. Complete, exact and efficient implementation for computing the adjacency graph of an arrangement of quadrics. In L. Arge, M. Hoffmann, and E. Welzl, eds., *15th Annual European Symposium on Algorithms*, Eilat, Israel, 2007, *LNCS 4698*, pp. 633–644. Springer.

[9] L. Dupont, D. Lazard, S. Lazard, and S. Petitjean. Near-optimal parameterization of the intersection of quadrics: I+II+III. *J. of Symbolic Computation*, 43(3):168–232, 2008.

- [10] A. Eigenwillig and M. Kerber. Exact and efficient 2d-arrangements of arbitrary algebraic curves. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA08)*, San Francisco, USA, 2008, pp. 122–131. ACM/SIAM.

- [11] A. Eigenwillig, M. Kerber, and N. Wolpert. Fast and exact geometric analysis of real algebraic plane curves. In C. W. Brown, ed., *Proceedings of the 2007 International Symposium on Symbolic and Algebraic Computation*, Waterloo, Ontario, Canada, 2007, pp. 151–158. ACM.

[12] A. Eigenwillig, L. Kettner, E. Schömer, and N. Wolpert. Exact, efficient and complete arrangement computation for cubic curves. *Computational Geometry*, 35(1-2):36–73, 2006.

- [13] P. Emeliyanenko. Visualization of points and segments of real algebraic plane curves. Masters thesis, Universität des Saarlandes, 2007.

- [14] P. Emeliyanenko and M. Kerber. Visualizing and exploring planar algebraic arrangements – a web application. In M. Teillaud and E. Welzl, eds., *Proceedings of the 24th ACM Symposium on Computational Geometry*, College Park Maryland, USA, 2008, pp. 224–225. ACM.

[15] P. Hachenberger and L. Kettner. 2d boolean operations on Nef polygons embedded on the sphere. In C. E. Board, ed., *CGAL User and Reference Manual*. 3.4 edition, 2008.

- [16] P. Hachenberger, L. Kettner, and K. Mehlhorn. Boolean operations on 3d selective Nef complexes: Data structure, algorithms, optimized implementation and experiments. *Computational Geometry: Theory and Applications*, 38(1-2):64–99, 2007.

[17] D. Halperin. Arrangements. In J. Goodman and J. O'Rourke, eds., *Handbook of Discrete and Computational Geometry*, ch. 24. CRC Press, 2nd edition, 2004.

- [18] M. Hemmer. *Exact Computation of the Adjacency Graph of an Arrangement of Quadrics*. Phd thesis, Johannes Gutenberg-Universität Mainz, Saarstr. 21, D 55122 Mainz, 2008.

- [19] M. Hemmer, S. Limbach, and E. Schömer. Continued work on the computation of an exact arrangement of quadrics. In *25th European Workshop on Computational Geometry*, Brussels, Belgium, 2009, pp. 313–316. Université Libre de Bruxelles.

[20] S. Lazard, L. M. Peñaranda, and S. Petitjean. Intersecting quadrics: an efficient and exact implementation. In *SCG '04: Proceedings of the twentieth annual symposium on Computational geometry*, New York, NY, USA, 2004, pp. 419–428. ACM.

[21] F. Messine. Extension of affine arithmetic: Application to unconstrained global optimization. *Journal of Universal Computer Science*, 8:992–1015, 2002.

[22] M. Meyerovitch. Robust, generic and efficient construction of envelopes of surfaces in three-dimensional spaces. In Y. Azar and T. Erlebach, eds., *Proceedings of 14th Annual European Symposium on Algorithms (ESA)*, 2006, *LNCS 4168*, pp. 792–803. Springer.

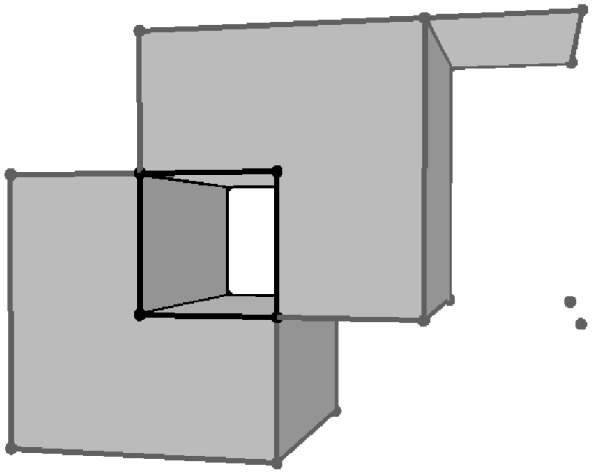


Figure 28.8: A Nef polyhedron with non-manifold edges, a dangling facet, and two isolated vertices. The tunnel boundary does not belong to this point set.

Boolean Operations on 3D Selective Nef Complexes

Investigators: P. Hachenberger (now University Eindhoven), Lutz Kettner (now Single Images), and Kurt Mehlhorn

Nef polyhedra in d-dimensional space are the closure of half-spaces under boolean set operations. In consequence, they can represent non-manifold situations, open and closed sets, mixed-dimensional complexes, and they are closed under all boolean and topological operations, such as complement and boundary. They were introduced by W. Nef in his seminal 1978 book on polyhedra. The generality of Nef complexes is essential for some applications.

In [1], we present a new data structure for the boundary representation of three-dimensional Nef polyhedra and efficient algorithms for boolean operations. We use exact arithmetic to avoid well known problems with floating-point arithmetic and handle all degeneracies. Furthermore, we present important optimizations for the algorithms, and evaluate this optimized implementation with extensive experiments. The experiments supplement the theoretical runtime analysis and illustrate the effectiveness of our optimizations. We compare our implementation with the Acis CAD kernel. Acis is mostly faster, by a factor up to six. There are examples on which Acis fails.

The implementation was released as Open Source in the Computational Geometry Algorithm Library (Cgal) release 3.1 in December 2004.

References

- [1] P. Hachenberger, L. Kettner, and K. Mehlhorn. Boolean operations on 3d selective Nef complexes: Data structure, algorithms, optimized implementation and experiments. *Computational Geometry: Theory and Applications*, 38(1-2):64–99, 2007.

28.7.4 Analysis and Topology Computation of Surfaces

In the past decades, the computation of the topology of real algebraic curves and surfaces has received a lot of attention in algebraic geometry, computer graphics and computer aided geometric design. Beside the theoretical interest of the problem, accurate topological and geometric information of algebraic objects is crucial for a good visualization and for a meaningful approximation by simpler objects, such as splines or polygons. What do we mean by the topological analysis of an algebraic surface S? Here we mainly distinguish two different descriptions. The first, based on cylindrical algebraic decomposition (cad for short), outputs a *stratification* of S, that is, a decomposition Ω_S of S into nonsingular, connected sets, such that the boundary of each component is the union of other components in Ω_S. Ω_S is obtained by a decomposition of the plane into 0, 1 and 2-dimensional connected sets, which are *delineable*. We call a connected set delineable if, with respect to some projection direction (usually with respect to the $z-$coordinate), its fiber consists of a certain number of disjoint function graphs. In this setting the topological analysis constitute the decomposition Ω_S of S into these function graphs together with the information how they are connected to each other. Another, and even more crucial, method to encode the topology of S is to determine a simplicial complex $\mathcal{T}_S$ that is isotopic to the surface S (see Figure 28.9 for an example). Figuratively speaking, this means that it is possible to continuously deform $\mathcal{T}_S$ into S without causing any intersections during the deformation process. In the previous reporting period we contribute major results with respect to this subject. Based on our work on arrangements of algebraic plane curves, we have been the first (and so far only) that designed and implemented an efficient method for computing the topology of algebraic surfaces. This includes a general software framework to handle a set of (semi-)algebraic surfaces and a concrete instantiation for one algebraic surface. Herein the stratification algorithm and, on top of that, a triangulation is implemented.

A Framework for the Geometrical and Topological Analysis of (Algebraic) Surfaces

Investigators: Eric Berberich and Michael Sagraloff

Let $\mathcal{S}$ in $\mathbb{R}^3$ be a set of (semi-)algebraic surfaces. In order to capture the geometric and topological information given by $\mathcal{S}$, we developed a generic framework that abstracts combinatorial and topological tasks from geometric operations [3, 4]. It aims at supporting various algorithms and applications on surfaces in computational geometry. Our implementation follows the generic programming paradigm, that is, to support a certain family of surfaces, we require a small set of types and some basic operations on them. All such are collected in a class that has to implement the newly presented SurfaceTraits_3 concept. Its design is simple, the interface intuitive, and we do not assume generic position.

The framework decomposes into two parts that are also reflected by the geometric operations expected from surfaces. First, important 0- and 1-dimensional features of a surface (or pairs

of them) are projected onto the xy-plane, obtaining a finite planar arrangement $\mathcal{A}_\mathcal{S}$ with certain properties, that is, each component is delineable with respect to each surface $S \in \mathcal{S}$. For this projection step we rely on CGAL's `Arrangement_2` package as basic tool. The points of a cell of $\mathcal{A}_\mathcal{S}$ share some invariant properties. In particular, they have the same z-pattern. A z-pattern at a point p encodes the sequence of intersections of $S_i \in \mathcal{S}$ with the vertical line ℓ_p at p. It then suffices to compute a z-pattern only for a sample point of each cell of $\mathcal{A}_\mathcal{S}$ to lift the planar cells. We also compute their adjacency relation and their position in $\mathbb{R}^3$. We go into this direction as various algorithms in computational geometry can be expressed in terms of this information, especially when expressed with respect to a small number of surfaces involved at once. Examples of such applications are meshing of single surfaces, the computation of space-curves defined by two surfaces, computing lower envelopes of surfaces, or supporting the computation of a three-dimensional arrangement with basic predicates. Space curves and lower envelopes are already implemented, the others are open for future research.

The framework constitutes a *complete* tool for the study of surfaces that are usually considered in CAD applications, as we have also shown, that the well-known family of (semi-) algebraic surfaces fulfills the framework's requirements; see again [3]. This includes simple algebraic surfaces as for instance planes, spheres, ellipsoids, or even more difficult surfaces as given by polyhedral surfaces or B-spline surfaces. So far, robust implementations on these surfaces were not available. We consider the framework to be an important step to fill this gap. In particular, we can also instantiate the framework by a fully-fledged model for special algebraic surfaces, namely quadrics having degree 2.

We also developed an instantiation for the framework that supports a single algebraic surface, but of arbitrary degree. Details on their stratification and triangulation based on the obtained lifting are presented below. Future research aims at implementing an efficient full model supporting also pairs of algebraic surfaces, which enables further operations. Another goal is provide concretizations for surfaces being piecewise algebraic or for B-splines. On the combinatorial side, a main goal is to turn the framework as a central tool to support the construction of three-dimensional arrangements.

Stratification and Triangulation of an Algebraic Surface

Investigators: Eric Berberich, Pavel Emeliyanenko, Michael Kerber, and Michael Sagraloff

As already mentioned at beginning of this section we developed an instantiation [1, 2] for the previously described framework that supports a single algebraic surface. We provide a complete C++-implementation that computes the exact topology of a real algebraic surface S, implicitly given by a polynomial $f \in \mathbb{Q}[x, y, z]$ of arbitrary total degree N. Additionally, our analysis provides geometric information as it supports the computation of arbitrary precise samples of S including critical points. As proposed by the framework the method follows a projection approach. A planar arrangement is constructed such that all cells are delineable with respect to the surface S. This induces a stratification Ω_S of S into $O(N^5)$ nonsingular cells. The analysis also comprises the complete adjacency information between these cells. Furthermore, the proposed approach applies numerical and combinatorial methods to minimize costly symbolic computations. The algorithm handles all kinds of degeneracies without transforming the surface into a generic position. Based on Ω_S our implementation also

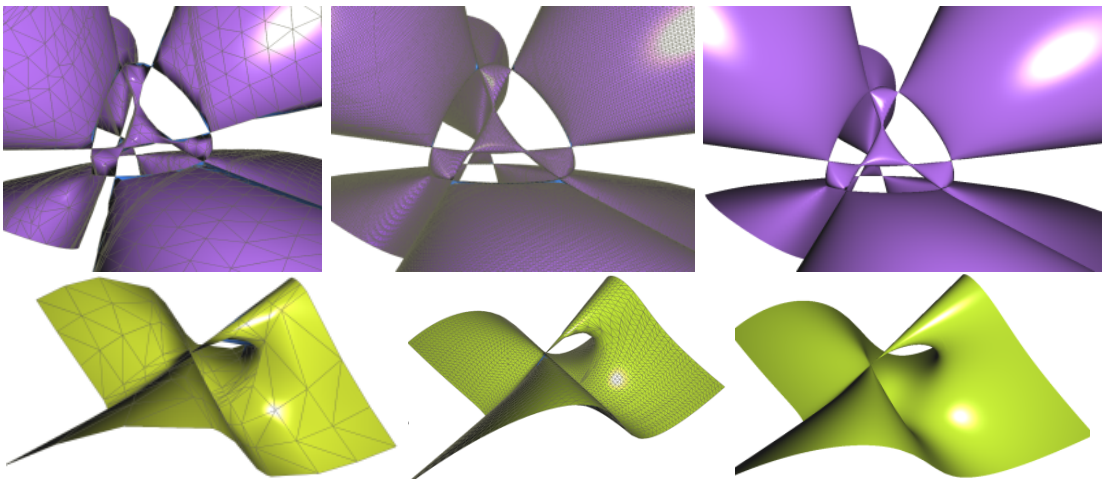


Figure 28.9: **First row:** *Kummer* surface of total degree 4. Mesh on the left has about 6000 triangles, in the middle – 200000 triangles, on the right -0 the same mesh without triangles shown. **Second row:** *KM42* surface of total degree 3. Mesh on the left has about 270 triangles, in the middle – 6800 triangles, on the right – the same mesh without triangles shown.

computes a simplicial complex which is isotopic to S. Whereas the basic triangulation consists of $O(N^7)$ triangles our software also meets the demand for arbitrary fine triangulations which, for instance, are asked for visualization purposes (see Figure 28.9). Our work constitutes the first complete implementation to compute a stratification and a triangulation of algebraic surfaces. It already shows good performance for many well-known examples from algebraic geometry. In future work we plan to extend our results toward arrangements of algebraic surfaces. Analogous to the arrangement computation of plane algebraic curves we want to make our results accessible to the general public, that is, we plan to integrate the surface analysis into our Webdemo, presented in Section 28.7.3).

References

- [1] E. Berberich, M. Kerber, and M. Sagraloff. Exact geometric-topological analysis of algebraic surfaces. In M. Teillaud and E. Welzl, eds., *Proceedings of the 24th ACM Symposium on Computational Geometry*, College Park Maryland, USA, 2008, pp. 164–173. ACM.

- [2] E. Berberich, M. Kerber, and M. Sagraloff. An efficient algorithm for the stratification and triangulation of an algebraic surface. Accepted for Computational Geometry: Theory and Applications - SoCG'08 special issue, 2009.

- [3] E. Berberich and M. Sagraloff. A generic and flexible framework for the geometrical and topological analysis of (algebraic) surfaces. In E. Haines and M. McGuire, eds., *Proceedings of the 2008 ACM Symposium on Solid and Physical Modeling*, Stony Brook, USA, 2008, pp. 171–182. ACM.

- [4] E. Berberich and M. Sagraloff. A generic and flexible framework for the geometrical and topological analysis of (algebraic) surfaces. Accepted for Computer Aided Geometric Design, 2009.

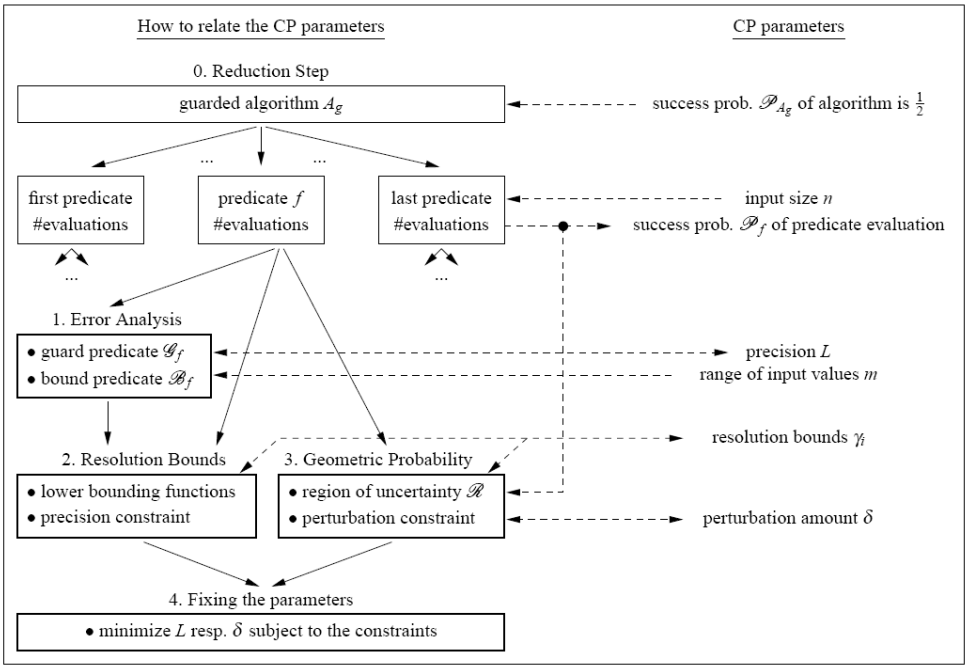


Figure 28.10: Flow chart of the analysis and the connection between the CP parameters.

28.7.5 Controlled Perturbation

Investigators: Kurt Mehlhorn, Ralf Osbild, and Michael Sagraloff

Controlled Perturbation (CP) constitutes an approach orthogonal to EGC. EGC insists on computing the exact result, CP is satisfied with an approximation. Whereas EGC algorithms must handle all degeneracies, CP handles no degeneracies.

The idea is as follows: Given a finite collection $\mathcal{C}$ of geometric objects (e.g. a collection of circles in the plane) and a structure $S(\mathcal{C})$ defined on $\mathcal{C}$ (e.g. the arrangement of the circles), CP considers a slight perturbation $\mathcal{C}'$ of $\mathcal{C}$ such that $S(\mathcal{C}')$ is degeneracy free and all predicate evaluations in algorithms to compute $S(\mathcal{C}')$ can be evaluated by fixed precision arithmetic. Starting with a certain, maximal perturbation δ amount and a precision L, the algorithm tries to compute $S(\mathcal{C}')$ with arithmetic of precision L. If some decision cannot be made safely at the current precision, δ and/or L are increased and a new perturbation $\mathcal{C}'$ is chosen. Then the algorithm starts over again.

In the previous reporting period, we developed a formal and general theory for CP [1] which we refined in the reporting period [2]. It identifies a large class of geometric algorithms that fit into the CP framework. Furthermore, we showed which combinations of parameters δ and L guarantee probability 1/2 for a successful perturbation. We give such an analysis for all predicates that can be expressed as multivariate polynomials. This constitutes an enormous simplification in the study of CP algorithms compared to former approaches. Previously, each algorithm had to be checked individually for possible adaption to CP, including an analysis of the CP parameters. In contrast, our framework and general analysis applies to a large class of algorithms.

In the future, we will pursue the following directions. First, we will aim for an improved analysis of polynomial predicates using ideas from our work on root isolation. Second, most algorithms in classical computational geometry use operations on objects such as points, lines and circles. Thus, in order to analyze a CP version of geometric algorithms, there is a need to also consider predicates, resulting from compass and ruler constructions. In general, these predicates cannot be expressed as multi-variate polynomials. Finally, we want to investigate whether CP is applicable to root isolation or topology computations.

References

[1] K. Mehlhorn, R. Osbild, and M. Sagraloff. Reliable and efficient computational geometry via controlled perturbation. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds., *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Part I*, Venice, Italy, 2006, *LNCS 4051*, pp. 299–310. Springer.

- [2] K. Mehlhorn, R. Osbild, and M. Sagraloff. A general approach to the analysis of controlled perturbation algorithms. Technical Report ACS-TR-361502-02, University of Groningen, 9700 AB Groningen THE NETHERLANDS, 2008.

28.8 Combinatorial Optimization

Coordinators: Ernst Althaus, Khaled Elbassioni, and Kurt Mehlhorn

Many real world applications are naturally formulated as combinatorial optimization problems, i.e. problems of finding the best solution(s) out of a finite set. Various methods have been developed to tackle such problems: integer programming, approximation algorithms and combinatorial algorithms, among others. D1 worked on applying these methods to various problems from different areas, ranging from bioinformatics to geometry, to scheduling, to mathematical programming, and several other areas. We can generally divide our work in this area into four parts: approximation algorithms, algorithm engineering, combinatorial algorithms, and computational issues related to transversal hypergraphs and polyhedra.

A good number of researchers in the group are working on developing approximation algorithms for a number of hard optimization problems, where the aim is to design either algorithms with improved guarantees or ones that work in a more general setting. Section 28.8.1 summarizes these results. In Section 28.8.2, we give an overview of the work on algorithm engineering for solving optimization problems, where the emphasis is on developing and evaluating algorithms for problems that arise directly from real-life applications. We report in Section 28.8.3 about our results on developing combinatorial algorithms that are either the first polynomial time algorithms for the problem considered, or improve the previous running time complexity.

While the results in Section 28.8.4 do not fall directly under Optimization, they are very much related. In particular, the section describes results on testing polyhedral properties, such as being integral or half-integral, and on finding the vertices of polyhedra and related problems. Clearly, the study of such notions is central to the use of linear programming to develop approximation algorithms.

28.8.1 Approximation Algorithms for Hard Optimization Problems

One of the most basic methods to deal with NP-hard optimization problems is to design polynomial-time algorithms that find a solution which is provably not far from the optimum. As we will see, our work spans different areas, including graph problems, computational economics, computational geometry, metric embeddings, mathematical programming, and scheduling.

Small Hop-diameter Sparse Spanners for Doubling Metrics

Investigator: Hubert Chan

The study of finite metrics and their properties has been a very fruitful area of research, with applications to many different problems. Many commonly arising problems (e.g., clustering, near-neighbor finding, network routing, just to name a few) deal with sets of points on which a distance function has been defined, and one wants to store and process this metric in different ways.

We focus on obtaining sparse representations of metrics: these are called *spanners*, and they have been studied extensively for both general and Euclidean metrics. Formally, a t-spanner for a metric $M = (V, d)$ is a weighted undirected graph $G = (V, E)$ such that the distances according to d_G (the shortest-path metric of G) are close to the distances in d: specifically, $d(u, v) \leq d_G(u, v) \leq t\, d(u, v)$. Clearly, one can take a complete graph and obtain $t = 1$, and hence the quality of the spanner is typically measured by how few edges can G contain whilst maintaining a *stretch* of at most t. The notion of spanners has been widely studied for general metrics (see, e.g. [1, 6, 8]), and for geometric distances (see, e.g., [2, 3, 9, 10]).

Very recently, there have been good constructions of spanners for doubling metrics as well: given a metric with doubling dimension[7] dim, the results of Chan et al. [4], and independently, those of Har-Peled and Mendel [7] show how to construct $(1+\varepsilon)$-spanners with $n(1+1/\varepsilon)^{O(\dim)}$ edges, where $n = |V|$ is the number of points in the metric.

In the paper [5], we extend these results to find spanners that also have small *hop diameter*. A t-spanner has hop diameter D if every pair $u, v \in V$ are connected by some path in G having length at most $t\, d(u, v)$, and furthermore there are at most D edges on this path. We show that given any metric with constant doubling dimension k, and any $0 < \varepsilon < 1$, one can find a $(1 + \varepsilon)$-spanner for the metric with a nearly linear number of edges (i.e., only $O(n \log^* n + n\varepsilon^{-O(k)})$ edges) and *constant* hop diameter; we can also obtain a $(1+\varepsilon)$-spanner with a linear number of edges (i.e., only $n\varepsilon^{-O(k)}$ edges) that achieves a hop diameter that grows like the functional inverse of Ackermann's function. Moreover, we prove that such tradeoffs between the number of edges and the hop diameter are asymptotically optimal.

References

[1] I. Althöfer, G. Das, D. Dobkin, D. Joseph, and J. Soares. On sparse spanners of weighted graphs. *Discrete Comput. Geom.*, 9(1):81–100, 1993.

[7] that is, one in which every ball can be covered by a constant number of balls of half the radius

[2] S. Arya, G. Das, D. M. Mount, J. S. Salowe, and M. H. M. Smid. Euclidean spanners: short, thin, and lanky. In *Proceedings of the 27th annual ACM symposium on Theory of computing (STOC)*, New York, NY, USA, 1995, pp. 489–498. ACM.

[3] P. B. Callahan and S. R. Kosaraju. A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields. *J. Assoc. Comput. Mach.*, 42(1):67–90, 1995.

[4] H. T.-H. Chan, A. Gupta, B. M. Maggs, and S. Zhou. On hierarchical routing in doubling metrics. In *Proceedings of the 16th annual ACM-SIAM symposium on Discrete algorithms (SODA)*, Philadelphia, PA, USA, 2005, pp. 762–771. Society for Industrial and Applied Mathematics.

• [5] T.-H. H. Chan. Small hop-diameter sparse spanners for doubling metrics. *Discrete and Computational Geometry*, 41(1):28–44, 2009.

[6] B. Chandra, G. Das, G. Narasimhan, and J. Soares. New sparseness results on graph spanners. *Internat. J. Comput. Geom. Appl.*, 5(1-2):125–144, 1995. Eighth Annual ACM Symposium on Computational Geometry (Berlin, 1992).

[7] S. Har-Peled and M. Mendel. Fast construction of nets in low dimensional metrics, and their applications. *Symposium on Computational Geometry*, pp. 150–158, 2005.

[8] D. Peleg and A. A. Schäffer. Graph spanners. In *J. Graph Theory 13*, 1989, pp. 99–116.

[9] J. S. Salowe. Constructing multidimensional spanner graphs. In *Internat. J. Comput. Geom. Appl. 1*, 1991, pp. 99–107.

[10] P. M. Vaidya. A sparse graph almost as good as the complete graph on points in k dimensions. In *Discrete Comput. Geom. 6*, 1991, pp. 369–381.

A QPTAS for TSP with Fat Weakly Disjoint Neighborhoods in Doubling Metrics

Investigators: Hubert Chan and Khaled Elbassioni

We consider the Traveling Salesman Problem with Neighborhoods (TSPN) in a metric space (V, d). An instance of the problem is given by a collection W of n subsets $\{P_1, P_2, \ldots, P_n\}$ in V. Each subset $P_j \subset V$ is known as a *neighborhood* or *region*. The objective is to find a minimum length tour that visits at least one point from each region.

This problem generalizes the well-known Traveling Salesman Problem (TSP), for which there are PTAS's for low-dimensional Euclidean metrics [8, 2, 10], and a QPTAS for doubling metrics [11]. The neighborhood version of the problem was first introduced by Arkin and Hassin [1], who gave constant approximation for the case when the regions are in the plane and "well-behaved" (e.g., disks, parallel and similar length segments, bounded ratio between the largest and smallest diameters). The general version of the problem was shown to have an inapproximability threshold of $\Omega(\log^{2-\epsilon} n)$ for any $\epsilon > 0$ by Halperin and Krauthgamer [7]. There is an almost matching upper bound of $O(\log N \log k \log n)$-approximation, using the results of Garg et al. [6] and Fakcharoenphol et al. [5], where N is the total number of points in V and k is the maximum number of points in each region.

The best previously known result for getting a $(1 + \varepsilon)$-approximation is by Mitchell [9], who obtained a PTAS for the Euclidean plane, where the regions are fat and almost disjoint. This result is obtained by the "guillotine subdivision" technique, which unfortunately only works for 2 dimensions. On the other hand, the hierarchical decomposition technique by

Arora [2] and Talwar [11] is applicable to more general metrics. However, as pointed out by Mitchell [9], previous attempts in applying this technique have led to only limited success.

In [3], we obtain some partial results. In particular, we give a $(1 + \varepsilon)$-approximation for instances on metrics with bounded doubling dimension. This includes low-dimensional Euclidean metrics, and hence is a generalization of Mitchell's result [9] for 3 or more dimensions. Moreover, since the doubling dimension is well defined for any metric, our framework covers metrics that do not have any geometric structure, and the regions need not be convex or even connected, where such notions might not even be applicable in the first place.

Nevertheless, we still need to place some restrictions on the regions, because the problem is APX-hard in general on the plane [4], which has bounded doubling dimension. We combine the notions of diameter variation, fatness and disjointness for geometric spaces, and define for regions in general metrics the notion of *α-fat weak disjointness*. We assume that the regions have a bounded number Δ types of radii. For the regions within the same type, there is some $\rho > 0$ such that there is a ρ-packing[8] consisting of one point from each region, and all the regions have diameters at least ρ and at most $O(\alpha\rho)$.

Our definition allows very general regions. Intuitively, all we require is that regions of similar diameters should each designate a point within, such that these points are far away from one another; the regions can otherwise intersect arbitrarily. The assumption that there are only a bounded number Δ of types of region diameters is also necessary, as otherwise the problem remains APX-hard.[9] Of course, the catch with working on such weak assumptions is that the running time of our algorithm is only quasi-polynomial. This is not surprising, because there is only a QPTAS known even for TSP on doubling metrics by Talwar [11].

References

[1] E. M. Arkin and R. Hassin. Approximation algorithms for the geometric covering salesman problem. *Discrete Applied Mathematics*, 55(3):197–218, 1994.

[2] S. Arora. Approximation algorithms for geometric TSP. In *The traveling salesman problem and its variations*, *Comb. Optim.*, vol. 12, pp. 207–221. Kluwer Acad. Publ., Dordrecht, 2002.

[3] H. Chan and K. Elbassioni. A QPTAS for TSP with fat weakly disjoint neighborhoods in doubling metrics. Technical Report 2009-01, DIMACS Technical Reports, 2009.

[4] M. Dror and J. B. Orlin. Combinatorial optimization with explicit delineation of the ground set by a collection of subsets. In *SIAM J. Discrete Math. 21(4)*, 2008, pp. 1019–1034.

[5] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *Proceedings of the 35th ACM symposium on Theory of computing (STOC)*, 2003, pp. 448–455. ACM Press.

[6] N. Garg, G. Konjevod, and R. Ravi. A polylogarithmic approximation algorithm for the group Steiner tree problem. *Journal of Algorithms*, 37(1):66–84, 2000. (Preliminary version in *9th SODA*, pages 253–259, 1998).

[7] E. Halperin and R. Krauthgamer. Polylogarithmic inapproximability. In *Proceedings of the thirty-fifth ACM symposium on Theory of computing*, 2003, pp. 585–594. ACM Press.

[8] A ρ-packing is a set of points with inter-point distance larger than ρ.

[9] However, this assumption is not necessary in the case of Euclidean metric.

[8] J. S. B. Mitchell. Guillotine subdivisions approximate polygonal subdivisions: a simple polynomial-time approximation scheme for geometric TSP, k-MST, and related problems. *SIAM J. Comput.*, 28(4):1298–1309, 1999.

[9] J. S. B. Mitchell. A PTAS for TSP with neighborhoods among fat regions in the plane. In N. Bansal, K. Pruhs, and C. Stein, eds., *Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms (SODA)*, 2007, pp. 11–18. SIAM.

[10] S. B. Rao and W. D. Smith. Approximating geometrical graphs via "spanners" and "banyans". In *Proceedings on 30th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 540–550. ACM, New York, 1998.

[11] K. Talwar. Bypassing the embedding: algorithms for low dimensional metrics. In *Proceedings of the 36th annual ACM symposium on Theory of computing (STOC)*, New York, NY, USA, 2004, pp. 281–290. ACM.

Metric Embeddings with Relaxed Guarantees

Investigator: Hubert Chan

Over the past decade, the field of metric embeddings has gained much importance in algorithm design. The central genre of problem in this area is the mapping of a given metric space into a "simpler" one, in such a way that the distances between points do not change too much. More formally, an *embedding* of a finite metric space (V, d) into a *target* metric space (V', d') is a map $\varphi : V \to V'$. Recent work on embeddings has used *distortion* as the fundamental measure of quality; the distortion of an embedding is the worst multiplicative factor by which distances are increased by the embedding[10]. The popularity of distortion has been driven by its applicability to approximation algorithms: if the embedding $\varphi : V \to V'$ has a distortion of D, then the cost of solutions to some optimization problems on (V, d) and on $(\varphi(V), d')$ can only differ by some function of D; this idea has led to numerous approximation algorithms [3].

In the context of some networking applications, however, distortion as defined above has turned out to be too demanding an objective function. Instead, the recent networking work has provided *empirical* guarantees of the following form: if we allow a small fraction of all distances to be *arbitrarily* distorted, we can embed the remainder with constant distortion in constant-dimensional Euclidean space. Such guarantees are natural for the underlying networking applications; essentially, a very small fraction of the location-based lookups may yield poor performance (due to the arbitrary distortion), but for the rest the quality of the embedding will be very good.

These types of results form a suggestive contrast with the theoretical work on embeddings. In particular, are the strong empirical guarantees for Internet latencies the result of fortuitous artifacts of this particular set of distances, or is something more general going on? To address this, Kleinberg, Slivkins, and Wexler [5] defined the notion of embeddings with *slack*: in addition to the metrics (V, d) and (V', d') in the initial formulation above, we are also given a *slack parameter* ε, and we want to find a map φ whose distortion is bounded by some quantity $D(\varepsilon)$ on all but an ε fraction of the pairs of points in $V \times V$. (Note that we allow the

[10] Formally, for an embedding $\varphi : V \to V'$, the *distortion* is the smallest D so that $\exists\, \alpha, \beta \geq 1$ with $\alpha \cdot \beta \leq D$ such that $\frac{1}{\alpha}\, d(x, y) \leq d'(\varphi(x), \varphi(y)) \leq \beta\, d(x, y)$ for all pairs $x, y \in V \times V$. Note that this definition of distortion is slightly non-standard—since $\alpha, \beta \geq 1$, it is no longer invariant under arbitrary scaling; however, this is merely for notational convenience, and all our results can be cast in the usual definitions of distortion.

distortion on the remaining εn^2 pairs of points to be arbitrarily large.) Roughly, Kleinberg et. al. [5] showed that any metric of bounded doubling dimension can be embedded with constant distortion into constant-dimensional Euclidean space, allowing a constant slack ε. Such metrics, which have been extensively studied in their own right, have also been proposed on several occasions as candidates for tractable abstractions of the set of Internet latencies (see e.g. [2, 4, 6, 7]).

In the paper [1], we answer some of the open questions posed in [5]. In particular, we show that provable guarantees of this type can in fact be achieved in general: any finite metric can be embedded, with constant slack and constant distortion, into constant-dimensional Euclidean space. We then show that there exist stronger embeddings into ℓ_1 which exhibit *gracefully degrading* distortion: there is a single embedding into ℓ_1 that achieves distortion at most $O(\log \frac{1}{\varepsilon})$ on all but at most an ε fraction of distances, *simultaneously* for all $\varepsilon > 0$. We extend this with distortion $O(\log \frac{1}{\varepsilon})^{1/p}$ to maps into general ℓ_p, $p \geq 1$ for several classes of metrics, including those with bounded doubling dimension and those arising from the shortest-path metric of a graph with an excluded minor. Finally, we show that many of our constructions are tight, and give a general technique to obtain lower bounds for ε-slack embeddings from lower bounds for low-distortion embeddings.

References

- [1] T.-H. H. Chan. Metric embeddings with relaxed guarantees. *SIAM Journal on Computing*, 38:2303–2329, 2009.

[2] M. Fomenkov, k. claffy, B. Huffaker, and D. Moore. Macroscopic Internet topology and performance measurements from the DNS root name servers. In *Usenix LISA*, 2001.

[3] P. Indyk. Algorithmic aspects of geometric embeddings. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2001.

[4] D. R. Karger and M. Ruhl. Finding nearest neighbors in growth-restricted metrics. In *Proceedings of the 34th Annual ACM Symposium on the Theory of Computing*, 2002, pp. 63–66.

[5] J. Kleinberg, A. Slivkins, and T. Wexler. Triangulation and embedding using small sets of beacons. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2004.

[6] T. Ng and H. Zhang. Predicting Internet network distance with coordinates-based approaches. In *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, 2002.

[7] C. G. Plaxton, R. Rajaraman, and A. W. Richa. Accessing nearby copies of replicated objects in a distributed environment. *Theory Comput. Syst.*, 32(3), 1999.

1.5D Terrain Guarding

Investigators: Khaled Elbassioni and Julián Mestre in cooperation with Erik Krohn, Domagoj Matijević, and Domagoj Ševerdija

In the *1.5D terrain guarding* problem we are given a polygonal region in the plane determined by an x-monotone polygonal chain, and the objective is to find the minimum number of guards to place on the chain such that every point in the polygonal region is guarded. This

kind of guarding problems and its generalizations to 3-dimensions are motivated by optimal placement of antennas for communication networks; for more details see [2, 1] and the references therein.

In [4] we presented a 4-approximation algorithm for the problem of placing the fewest guards on a 1.5D terrain so that every point of the terrain is seen by at least one guard. This improves on the currently best approximation factor of 5 [3]. Unlike most of the previous techniques, our method is based on rounding the linear programming relaxation of the corresponding covering problem. Besides the simplicity of the analysis, which mainly relies on decomposing the constraint matrix of the LP into totally balanced matrices, our algorithm, unlike previous work, generalizes to the weighted and partial versions of the basic problem.

References

[1] B. Ben-Moshe, M. J. Katz, and J. S. B. Mitchell. A constant-factor approximation algorithm for optimal 1.5D terrain guarding. *SIAM Journal on Computing*, 36(6):1631–1647, 2007.

[2] D. Z. Chen, V. Estivill-Castro, and J. Urrutia. Optimal guarding of polygons and monotone chains. In *Proceedings of the 7th Canadian Conference on Computational Geometry*, 1995, pp. 133–138.

[3] J. King. A 4-approximation algorithm for guarding 1.5-dimensional terrains. In *Proceedings of the 13th Latin American Symposium on Theoretical Informatics*, 2006, pp. 629–640.

• [4] E. Krohn, K. Elbassioni, D. Matijevic, J. Mestre, and D. Severdija. Improved approximation algorithms for 1.5D terrain guarding. In *26th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Freiburg - Germany, 2009. Internationales Begegnungs und Forschungszentrum f"ur Informatik (IBFI).

Scheduling to Minimize Lateness

Investigator: Julián Mestre in cooperation with Samir Khuller

In [1] we re-examined the classical problem of minimizing maximum lateness which is defined as follows: given a collection of n jobs with processing times and due dates, in what order should they be processed on a single machine to minimize maximum lateness? The lateness of a job is defined as its completion time minus its due date. This problem can be solved easily by ordering the jobs in non-decreasing due date order.

Consider the following question: which subset of k jobs should we reject to reduce the maximum lateness by the largest amount? While this problem can be solved optimally in polynomial time, we showed the following surprising result: there is a fixed ordering of the jobs, such that for all k, if we reject the first k jobs from this ordering, we derive an optimal solution for the problem in which we are allowed to reject k jobs. This allows for an incremental solution in which we can keep incrementally rejecting jobs if we need a solution with lower maximum lateness value. Moreover, we also developed an optimal $O(n \log n)$ time algorithm to find this ordering.

References

• [1] S. Khuller and J. Mestre. An optimal incremental algorithm for minimizing lateness with rejection. In *16th Annual European Symposium on Algorithms (ESA)*, Karlsruhe, Germany, 2008, *LNCS 5193*, pp. 601–610. Springer.

Scheduling with Limited Machine Availability

Investigators: Nicole Megow in cooperation with Jose Verschae, Alberto Marchetti-Spaccamela, Martin Skutella, and Leen Stougie

In classical scheduling theory it is assumed that resources are available continuously throughout the entire planning period. This is hardly the case in practice since there are working shifts, planned maintenance periods, or even unexpected machine breakdowns. Therefore, research on scheduling with limitations in the resource availability is certainly of practical importance, which is also reflected by the large number of publications on this field.

We consider scheduling on a single machine with limited availability to minimize the sum of weighted completion times. In [2], we consider the special case in which there is only a single non-available time period. We provide a preemptive algorithm with an approximation ratio arbitrarily close to the Golden Ratio, $(1+\sqrt{5})/2+\epsilon$, which improves on a previously best known 2-approximation. The non-preemptive version of the same algorithm yields a $(2+\epsilon)$-approximation. In [1] we improve these results by presenting a fully polynomial-time approximation scheme for both problems.

In [1] we affirmatively answer the longstanding open question whether a constant approximation algorithm exists for the problem with arbitrary unavailability periods. Here, we consider the much more general model of full uncertainty about the breakdowns. We design a polynomial time deterministic algorithm that finds a robust prefixed scheduling sequence with a solution value within 4 times the value an optimal clairvoyant algorithm can achieve, knowing the disruptions in advance and even being allowed to interrupt jobs at any moment. A randomized version of this algorithm attains in expectation a ratio of e w.r.t. a clairvoyant optimum. We show that such a ratio can never be achieved by any deterministic algorithm by proving that the price of robustness of any such algorithm is at least $1+\sqrt{3} \approx 2.73205 > e$.

References

[1] A. Marchetti-Spaccamela, N. Megow, M. Skutella, and L. Stougie. Robust sequencing on a single machine. Submitted, 2009.

[2] N. Megow and J. Verschae. A note on scheduling on a single machine with one non-availability period. Submitted, 2008.

On Eulerian Extension Problems and their Application to Sequencing Problems

Investigators: Nicole Megow in cooperation with Wiebke Höhn and Tobias Jacobs

We present in [2] a new technique for analyzing sequencing problems such as the variant of the *Traveling Salesman Problem* (TSP) studied first by Gilmore and Gomory [1] and related flowshop scheduling problems. We show that those sequencing problems have a natural interpretation as *Eulerian Extension Problems* which leads to new structural insights and solution methods. On a high level view, for an instance of a sequencing problem we find a particular Eulerian graph in which all existing Eulerian circuits represent sequencing solutions with the same cost. In fact, we provide the entire set of optimal solutions, instead of just a single one.

Besides the well-known Gilmore-Gomory TSP and no-wait flowshop makespan scheduling, we also consider a non-standard flowshop sequencing problem. This problem concerns no-wait flowshop scheduling with the objective of minimizing the number of *interruptions*, i.e., the number of maximal idle time intervals on the last production stage. This problem is motivated by a particular application in steel production, the continuous casting process, in which ladles of melted steel have to pass several production stages. The final stage, the casting machine, plays a special role: the steel must flow continuously into the casting machine. When the flow is broken (we call it interruption), then the casting machine must be stopped for maintenance and extensive cleaning. Therefore, practitioners call it their objective to minimize the number of interruptions. We show how this problem with two production stages can be solved as an Eulerian Extension Problem with a particular cost function. To resolve the complexity status of all other problem versions, we define a natural generalization of the Eulerian Extension Problem where vertices correspond to points in the two-dimensional space. We show that this problem is NP-hard and prove its equivalence to the sequencing problem.

References

[1] P. C. Gilmore and R. E. Gomory. Sequencing a one state-variable machine: A solvable case of the traveling salesman problem. *Operations Research*, 12:655–679, 1964.

[2] W. Höhn, T. Jacobs, and N. Megow. On eulerian extension problems and their applications to sequencing problems. Submitted, 2009.

Shutdown and Turnaround Scheduling

Investigators: Nicole Megow in cooperation with Rolf Möhring and Jens Schulz

Large-scale maintenance in industrial plants requires the entire shutdown of production units for disassembly, comprehensive inspection and renewal. This so-called turnaround is an important process but causes high out-of-service cost. Therefore a good schedule for the turnaround has a high priority to the manufacturer. A good schedule is not simply a short schedule. The project execution can be speeded up at the expense of adding resources, mostly in the form of additional workers. Thus, short projects cause high resource cost whereas cheap projects take a long time. Moreover, in practice, task execution times typically involve uncertainty. Such uncertainty arises due to unforeseen repair jobs, and, naturally, a short schedule is less robust against unexpected repair jobs or processing delays than a schedule with long duration that offers more flexibility for rescheduling. Such considerations are fundamental in the decision process of a turnaround project manager.

In [1], we derive models and algorithms for turnaround scheduling that include different features such as time-cost tradeoff, precedence constraints, hiring external resources, resource leveling, different working shifts, and risk analysis. We propose a framework for decision support that consists of two phases. The first phase supports the manager in finding a good makespan for the turnaround. It computes an approximate project time-cost tradeoff curve together with a stochastic evaluation of the risk for meeting a particular makespan t. Our risk measures are the expected tardiness at time t and the probability of completing the turnaround within time t. In the second, detailed planning phase, we solve the actual scheduling optimization problem for the makespan chosen in the first phase heuristically and

compute a detailed schedule that respects all side constraints. Again, we complement this by computing upper bounds for the same two risk measures, but now for the detailed schedule.

Our experimental results show that our methods solve large real-world instances with 100,000 – 150,000 jobs from chemical manufacturing plants very fast and yield an excellent resource utilization. A comparison with solutions of a mixed integer program on smaller instances proves the high quality of the schedules that our algorithms produce within a few minutes. To the best of our knowledge, this is the first time that the turnaround scheduling problem is treated with this combination of optimization techniques. The interest of our cooperation partners (the management consulting company T.A. Cook and their customers at chemical manufacturing sites) in these methods has led to a commercial initiative to integrate them as a software tool into Microsoft Project.

References

[1] N. Megow, R. Möhring, and J. Schulz. Decision support and optimization in shutdown and turnaround scheduling. Submitted, 2009.

Subcoloring and Hypocoloring Interval Graphs

Investigators: Rajiv Raman in cooperation with Sriram Pemmaraju, Rajiv Gandhi, and Brad Greening

In [1], we study the *sub-coloring* and *hypo-coloring* problems on interval graphs. These problems have applications in job scheduling and distributed computing and can be used as "subroutines" for other combinatorial optimization problems. In the sub-coloring problem, given a graph G, we want to partition the vertices of G into minimum number of sub-color classes, where each sub-color class induces a union of disjoint cliques in G. In the hypo-coloring problem, given a graph G, and integral weights on vertices, we want to find a partition of the vertices of G into sub-color classes such that the sum of the weights of the heaviest cliques in each sub-color class is minimized. We present a "forbidden subgraph" characterization of graphs with sub-chromatic number k and use this to derive a a 3-approximation algorithm for sub-coloring interval graphs. For the hypo-coloring problem on interval graphs, we first show that it is NP-complete and then via a reduction to the max-coloring problem, show how to obtain an $O(\log n)$-approximation algorithm for it.

References

[1] R. C. Gandhi, B. G. Jr., S. Pemmaraju, and R. Raman. Subcoloring and hypocoloring interval graphs. 2009.

Approximation Algorithms for Packing Problems

Investigators: Rolf Harren and Rob van Stee

Orthogonal packing problems are natural multidimensional generalizations of the classical bin packing problem and knapsack problem and occur in many different settings. The input consists of a set $I = \{r_1, \ldots, r_n\}$ of d-dimensional rectangular items $r_i = (a_{i,1}, \ldots, a_{i,d})$ and a space Q. The task is to pack the items in an orthogonal, i.e., axis-parallel, and non-overlapping

manner in the given space. In the bin packing setting, the space Q consists of an unlimited number of equally-sized bins. The objective is to pack all items in the minimal number of bins. In the strip packing setting, the space Q is given by a strip of bounded basis and unlimited height and we aim to minimize the overall height needed to pack all items. In the knapsack packing setting, the given space Q is a single, usually unit sized bin and the items have associated profits p_i. The goal is to maximize the profit of a selection of items that can be packed into the bin. We study all three objectives in different dimensions d.

For the 2-dimensional bin packing problem we show a 2-approximation algorithm in [4]. We also consider the same problem where we allow to rotate the items by 90 degrees and present a 2-approximation algorithm for this case as well [5] (a preliminary version appeared in [3]). Both algorithm are best possible unless P = NP.

For the 3-dimensional knapsack problem we derive a variety of approximation algorithms in [1], i.e., a $(7+\varepsilon)$-approximation algorithm for the problem without rotations and a $(6+\varepsilon)$-approximation and a $(5+\varepsilon)$-approximation for the problem where rotations of 90 degrees are permitted around the z-axis and around all axes, respectively.

For the restriction to instances where all the items are squares, cubes or hypercubes we derive the following further results for the strip packing problem and the knapsack packing problem for arbitrary but fixed dimensions d [2]. Our algorithm for the d-dimensional hypercube knapsack problem has an approximation guarantee of $1 + 1/2^d + \varepsilon$–which has the interesting property that the approximation ratio improves with increasing dimension. For the d-dimensional hypercube strip packing problem we present an asymptotic PTAS and for the so-called d-dimensional hypercube knapsack packing with large resources we show an algorithm which outputs a solution of value at least $(1-\varepsilon)\mathrm{OPT}$ if the volume of the knapsack is much larger than the items.

References

- [1] F. Diedrich, R. Harren, K. Jansen, R. Thöle, and H. Thomas. Approximation algorithms for 3d orthogonal knapsack. *Journal of Science and Technology*, 23(5):749–762, 2008.
- [2] R. Harren. Approximation algorithms for orthogonal packing problems for hypercubes. *Theoretical Computer Science*, to appear, 2009.
- [3] R. Harren and R. van Stee. Packing rectangles into 2 opt bins using rotations. In J. Gudmundsson, ed., *11th Scandinavian Workshop on Algorithm Theory*, Gothenburg, Sweden, 2008, *LNCS 5124*, pp. 306–318. Springer.

  [4] R. Harren and R. van Stee. An absolute 2-approximation algorithm for two-dimensional bin packing, 2009.
- [5] R. Harren and R. van Stee. Absolute approximation ratios for packing rectangles into bins. *Journal of Scheduling*, 2009. To appear.

Langrangian Relaxation and Partial Cover

Investigator: Julián Mestre

Lagrangian relaxation has been used extensively in the design of approximation algorithms. At a very high level, the technique allows us to turn an α-approximation for the related problem Prize Collecting version of the covering problem at hand into an approximation for

the Partial Cover problem we are interested in solving. (In the partial version of a covering problem we are allowed to leave a prescribed number of elements uncovered, while in the prize collecting version we must pay a penalty for each element not covered.)

In [2], we showed a lower bound of $\frac{4\alpha}{3}$ on the approximation factor achievable for Partial Cover using Lagrangian relaxation when no assumption is made about the α-approximation; this matches the upper bound of Könneman et al. [1]. To overcome this obstacle, we then concentrated on a specific algorithm for a broad class of covering problems: Totally Balanced Cover. By carefully analyzing the algorithm's inner workings, we identified structural properties that allowed us to give an almost tight characterization of the integrality gap of the standard linear relaxation for Partial Totally Balanced Cover. This in turn implies improved approximation algorithms for a number of related problems.

References

[1] J. Könemann, O. Parekh, and D. Segev. A unified approach to approximating partial covering problems. In Y. Azar and T. Erlebach, eds., *ESA*, 2006, *LNCS 4168*, pp. 468–479. Springer.

• [2] J. Mestre. Lagrangian relaxation and partial cover (extended abstract). In *25th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Bordeaux, France, 2008, pp. 539–550. Internationales Begegnungs- und Forschungszentrum f'ur Informatik (IBFI), Schloss Dagstuhl, Germany.

Adaptive Local Ratio

Investigator: Julián Mestre

Local Ratio is a well-known paradigm [1] for designing approximation algorithms for combinatorial optimization problems. At a very high level, a local ratio algorithm first decomposes the input weight function w into a positive linear combination of simpler weight functions or *models*. Guided by this process a solution S is constructed such that S is α-approximate with respect to each model used in the decomposition. As a result, S is α-approximate under w as well.

These models usually have a very simple structure that remains "unchanged" throughout the execution of the algorithm. In [2], we showed that adaptively choosing a model from a richer spectrum of functions can lead to a better local ratio. Indeed, by turning the search for a good model into an optimization problem of its own, we get improved approximations for a data migration problem. We hope that these findings encourage the study of non-uniform updates for local-ratio or primal-dual algorithms; perhaps in some cases, as in our problem, this may help realize the full potential of these techniques.

References

[1] R. Bar-Yehuda, K. Bendel, A. Freund, and D. Rawitz. Local ratio: A unified framework for approximation algorithms. *ACM Computing Surveys*, 36(4):422–463, 2004.

• [2] J. Mestre. Adaptive local ratio. In *19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, San Francisco, USA, 2008, pp. 152–160. Society for Industrial and Applied Mathematics.

Profit-maximizing Pricing for the Highway and Tollbooth Problems

Investigators: Khaled Elbassioni and Rajiv Raman in cooperation with Saurabh Ray and René Sitters

Consider the problem of pricing the bandwidth along the links of a network such that the revenue obtained from customers interested in buying bandwidth along certain paths in the network is maximized. Suppose that each customer declares a set of paths she is interested in buying, and a maximum amount she is is willing to pay for each path. The network service provider's objective is to assign single prices to the links such that the total revenue from customers who can afford to purchase their paths is maximized. Recently, numerous papers have appeared on the computational complexity of such pricing problems.

A special case of this problem, where each customer is interested in purchasing only a single path (*single-minded*), and where there is no upper bound on the number of customers purchasing each link (*unlimited supply*) was studied by Guruswami et al. [5] under the name of *tollbooth problem*. The authors of [5] showed that the problem is already APX-hard when the network is restricted to be a tree, and also presented a polynomial time algorithm for the case when all paths start at a certain root of the tree. In [5], the authors also studied the *highway problem*, a further restriction where the tree is a path, and gave polynomial-time algorithms when either the budgets are bounded and integral, or all paths have a bounded length.

We make progress on several special cases of the problem in [3]. For the tollbooth problem, we give an $O(\log n)$-approximation which is an improvement over the previous $O(\log m)$-approximation, since $n \leq 3m$ can be always assumed. We also show that if all the paths are going towards a certain root, then a $(1-\epsilon)$-approximation can be obtained in quasi-polynomial time. This result extends our recently developed quasi-PTAS [4] for the highway problem, and uses essentially the same technique. However, there are a number of technical issues that have to be resolved for this technique to work on trees; most notably is the use of the Separator Theorem for trees, and the modification of the price-guessing strategy to allow only for *one-sided* guesses.

The existence of a quasi-PTAS for the highway problem indicates that a PTAS or even an FPTAS is still a possibility, since the problem was only known to be weakly NP-hard [2]. We show that the highway problem is indeed strongly NP-hard and hence admits no FPTAS unless P=NP.

Balcan et al. [1] considered a model in which some items can be priced below zero (in the form of a discount) so that the overall profit is maximized. They gave a 4-approximation for the uniform budgets case, and a quasi-PTAS for a special case in which there is an optimal pricing that has only a bounded number of negatively priced items. In [3] we show that the existence of a quasi-PTAS in the general case is highly unlikely, by showing that the problem is APX-hard.

References

[1] M.-F. Balcan, A. Blum, H. Chan, and M. Hajiaghayi. A theory of loss-leaders: Making money by pricing below cost. In X. Deng and F. C. Graham, eds., *In proceedings of the 3rd International Workshop on Internet and Network Economics (WINE)*, 2007, *LNCS 4858*, pp. 293–299. Springer.

[2] P. Briest and P. Krysta. Single-minded unlimited supply pricing on sparse instances. In *SODA '06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, New York, NY, USA, 2006, pp. 1093–1102. ACM Press.

- [3] K. Elbassioni, R. Raman, S. Ray, and R. Sitters. On profit-maximizing pricing for the highway and tollbooth problems. 2009.

- [4] K. Elbassioni, R. Sitters, and Y. Zhang. A quasi-ptas for profit-maximizing pricing on line graphs. In L. Arge, M. Hoffmann, and E. Welzl, eds., *Algorithms - ESA 2007, 15th Annual European Symposium, Eilat, Israel, October 8-10, 2007, Proceedings*, Eilat, Israel, 2007, *LNCS 4698*, pp. 451–462. Springer.

[5] V. Guruswami, J. D. Hartline, A. R. Karlin, D. Kempe, C. Kenyon, and F. McSherry. On profit-maximizing envy-free pricing. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, Philadelphia, PA, USA, 2005, pp. 1164–1173. Society for Industrial and Applied Mathematics.

Maximum Feasible Subsystems

Investigators: Khaled Elbassioni and Rajiv Raman in cooperation with Saurabh Ray and René Sitters

Consider the pricing problem in the previous section and the situation when the objective is to maximize the *customer satisfaction* but yet make reasonable profit. One way to model this is to find a pricing which maximizes the number (or total weight) of customers that are able to purchase their bundles. In this case, it makes sense to assume that each purchasing customer will have to pay some minimum amount, e.g. a fraction of her/his budget. This variant of the problem seems to be much harder in the general case. Indeed it can be modeled as a special instance of the following *maximum feasible subsystem problem* (MRFS) which is of independent interest. Given a system of linear constraints $\ell_i \leq a_i^T x \leq u_i$, where $a_i \in \{0,1\}^n$, and $\ell_i, u_i \in \mathbb{R}_+$, for $i = 1, \ldots, m$, we seek to find the largest subsystem for which there exists a feasible solution $x \geq 0$. In [1], we presented approximation algorithms and inapproximability results for this fundamental problem, and studied some of its important special cases. Our main conclusions are:

1. a sharp separation in the approximability between the case when $L = \max\{\ell_1, \cdots, \ell_m\}$ (assuming w.l.o.g. that the smallest positive ℓ_i is 1) is bounded above by a polynomial in n and m, and the case when it is not;

2. for the case where the constraint matrix has the consecutive ones property, we show a sharp separation in approximability between the case where we allow a violation of the upper bounds by at most a $(1+\epsilon)$ factor, for any fixed $\epsilon > 0$, and the case when no violations are allowed.

In more details, we show in in [1] that if L is bounded by a polynomial in n and m, then the problem cannot be approximated beyond a factor of $O(\log^{\mu} n)$, for some positive μ even when $\ell_i = u_i = 1\ \forall i = 1, \cdots m$, unless NP$\subseteq$BPTIME$(2^{n^{\epsilon}})$. If L is not bounded by a polynomial in n and m, then we show that MRFS cannot be approximated beyond a factor $O(n^{1/3-\epsilon})$, for some $\epsilon > 0$, unless NP=ZPP. Along the way, we also show that the induced matching

	Approximation		Inapproximability
	(α, β)	Running time	(α, β)
General 0/1-matrices	$(\log(nL/\epsilon), 1+\epsilon)$	$poly(n, m, \log L, \frac{1}{\epsilon})$	$(O(\log^{\mu} n), O(1))$
			$(O((\frac{\log L}{\log\log L})^{\frac{1}{3}-\epsilon}), O(1))$
Interval matrices	$(2\log n, 2)$	$poly(n, m)$	$(O(1), 1)$
	$(O(\log^2 n \log\log(nL/\epsilon)), 1+\epsilon)$	$poly(n, m, \log L, \frac{1}{\epsilon})$	
	$(1, 1+\epsilon)$	$(mL)^{O(\frac{\log L}{\epsilon}\log^2 m)}$	
	$(\sqrt{m}\log n, 1)$	$poly(n, m)$	

Table 28.1: Summary of positive and negative approximability results for MRFS with 0/1-matrices: $\mu \in (0, 1)$ is assumed to be some fixed constant, while ϵ is any arbitrary constant in $(0, 1)$. The inapproximability result (f, g) should be interpreted as follows: under strongly believed complexity assumptions, any algorithm yielding a solution with violation $\beta = g$ cannot give a better approximation factor than f.

problem in bipartite graphs cannot be approximated beyond a factor of $O(n^{1/3-\epsilon})$, for any $\epsilon > 0$ improving the APX-hardness result of Duckworth, et al. (2005).

On the positive side, we consider (α, β)-approximations, i.e., a subsystem with size at least $|\mathsf{OPT}|/\alpha$, which admits a solution violating the upper bounds by a factor of at most β. For the important special case of the MRFS problem when the constraint matrix has the consecutive ones property, we show that if we allow the upper bounds to be violated by at most a $(1+\epsilon)$ factor, for any $\epsilon > 0$, then approximation factor of $\alpha = 1$ can be obtained in quasi-polynomial time, when $L = poly(n, m)$ and hence, is unlikely to be APX-hard. However, if no violations are allowed, we show that the problem becomes APX-hard, even when the underlying graph of the constraint matrix is a clique! In the case when L is not bounded by a polynomial in n and m, we present an $(O(\log^2 n \log\log(nL)), 1+\epsilon)$ polynomial-time approximation algorithm. Table 28.1 summarizes our results. Note that in the general case, the upper and lower bounds are almost tight (up to constant factors in the exponent).

References

- [1] K. Elbassioni, R. Raman, S. Ray, and R. Sitters. On the approximability of the maximum feasible subsystem problem with 0/1-coefficients. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2009*, New York, NY, USA, 2009, pp. 1210–1219. SIAM.

A Primal-dual SDP Algorithm to Approximate the Lovàsz ϑ-Function

Investigators: Hubert Chan, Kevin Chang, and Rajiv Raman

The Lovàsz ϑ-function [3] on a graph $G = (V, E)$ can be defined as the maximum of the sum of the entries of a positive semidefinite matrix X, whose trace $Tr(X)$ equals 1, and $X_{ij} = 0$ whenever $\{i, j\} \in E$. This function appears as a subroutine for many algorithms on graph problems such as maximum independent set and maximum clique. In this work [2], we apply

Arora and Kale's primal-dual method for SDP [1] to design an algorithm to approximate the ϑ-function within an additive error of $\delta > 0$ which runs in time $O(\frac{\alpha^2 n^2}{\delta^2} \log n \cdot M_e)$, where $\alpha = \vartheta(G)$ and $M_e = O(n^3)$ is the time for a matrix exponentiation operation. It follows that on perfect graphs G, our primal-dual method computes $\vartheta(G)$ exactly in time $O(\alpha^2 n^5 \log n)$. Moreover, our techniques generalize to the *weighted* Lovàsz ϑ-function, and both the maximum independent set weight and the maximum clique weight for vertex weighted graphs.

References

[1] S. Arora and S. Kale. A combinatorial, primal-dual approach to semidefinite programs. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, 2007, pp. 227–236.

[2] H. Chan, K. Chang, and R. Raman. A primal-dual sdp algorithm to approximate the lovàsz ϑ-function. 2009.

[3] L. Lovász. On the shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25:1–7, 1979.

28.8.2 Algorithm Engineering for Combinatorial Optimization Problems

We are working on the development, implementation and experimental evaluation of algorithms for combinatorial optimization problems. We considered some real life applications, such as the power efficient addressing of OLED-displays, problems from bioinformatics, and the problem of efficiently computing safe bounds on the value of a linear program while avoiding the use for expensive rational arithmetic.

Algorithms for longer OLED lifetime

Investigator: Andreas Karrenbauer

Organic **L**ight **E**mitting **D**iodes (**OLED**s) have received growing attention recently as more and more commercial products are equipped with such displays. Though they have many advantages over current technology, such as Liquid Chrystal Display (LCD), only small-sized OLED displays have entered the mass market.

OLED displays are considered to be the displays of the future. The image and video displayed has a very high contrast and a viewing angle of nearly 180 degrees. It reacts within 10 microseconds, which is much faster than the human eye can catch and is therefore well suited for video applications. Moreover, the display is physically flexible.

There are two different OLED technologies called *active matrix* (AM) and *passive matrix* (PM). The former is more expensive but offers a longer lifetime than the latter. Their limited lifetime is one major reason why there are only small-sized displays on the mass market. For mobile phones or digital cameras, large state of the art OLED displays are either too expensive or suffer from insufficient lifetime.

What causes this short lifetime? Briefly, the reason is the high electrical current flowing through the diodes that occurs with the traditional addressing techniques. Though it seems that every pixel shines continuously with a certain brightness, rows of pixels are continually lit up one after another. This works at a sufficiently high frame rate since the perception of the human eye is the average intensity emitted by each diode. The problem is that this

row-by-row activation scheme causes long idle times of the diodes and extreme stress upon activation.

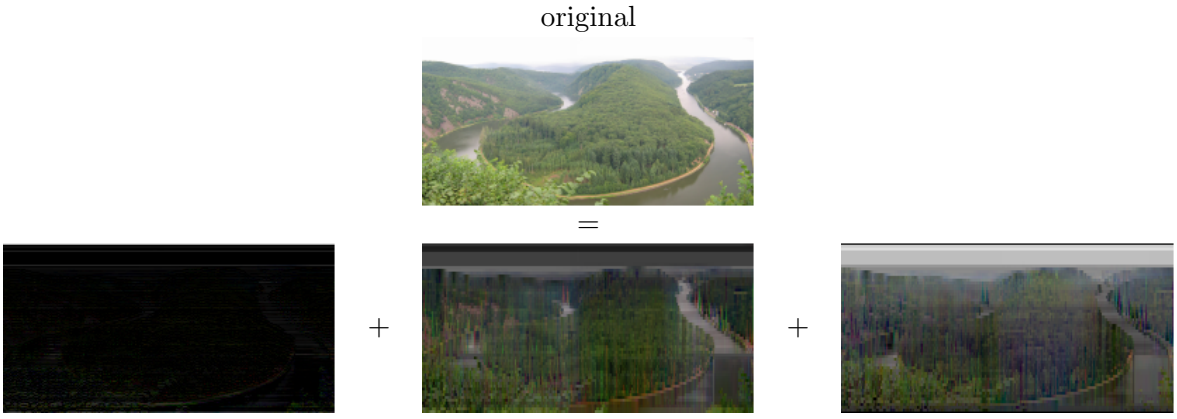


Figure 28.11: The original image on the top is decomposed into parts for CMLA2.

While a lot of research is conducted on the material science side, a joint project of the Max Planck Institute for Computer Science and the Electrical Engineering Department of Saarland University was started in mid-2005 to tackle this problem from the display driver point of view. As a first result of this cooperation, we filed a patent describing the **C**onsecutive **M**ulti**L**ine **A**ddressing (**CMLA**) scheme. It is a method to drive two or more consecutive rows simultaneously to lower the stress on the diodes. We will call the number k of rows that we maximally drive simultaneously the *multiline order* in the following. Occasionally, we will denote the CMLA driving scheme of a particular order k by CMLAk, e.g. CMLA2 for double-line addressing. The higher the multi-line order, the more we can potentially reduce the amplitude of the electrical current and, hence, increase the lifetime of passive matrix devices.

For such displays, rows can only be simultaneously displayed if their content is equal. Therefore the goal is to decompose an image into several subframes, the overlay (addition) of which is equal to the original image. Figure 28.11 shows such a decomposition. The first image (single) is traditionally displayed row-after-row. In the other two images (double) every two rows have the same content so that one can display these rows simultaneously. We thereby reduce the maximum intensities of the rows which directly relate to the necessary amplitude of the electrical current. Intuitively, the parts of the decomposition are much darker than the original one. The decomposition should be performed in such a way that the amplitude of the electrical current needed to display the picture is as small as possible. This involves a combinatorial optimization problem that must to be solved or approximated in real-time on a chip that drives such a device as shown in Figure 28.12.

We engineer an approximation algorithm that is part of Dialog Semiconductor's SmartXtend™technology, which is actually used to drive such an OLED display fabricated by TDK (see Figure 28.12). This imposes some restrictions on the methods and techniques we shall use. First of all, such an algorithm has to compute a feasible solution in real-time, i.e. faster than the perception of a human eye. Moreover, it should be implemented on a chip of low cost meaning that for example we are not able to use a general purpose LP solver or a general

Figure 28.12: This is a demonstrator fabricated by TDK using a display driver based on our algorithm.

purpose CPU with an IEEE floating point unit. We therefore look for algorithms that are sufficiently simple and easy to implement. Moreover, the algorithms should not suffer from numerical instabilities. That is, we want to compute a lossless decomposition. But exact rational arithmetic is not an option for such a real-time application. Rather, we want to use only fixed-precision number types, i.e. integers of a fixed size. Hence, we aim at a fully combinatorial algorithm using only addition, subtraction, and comparison.

Though this particular optimization problem with limited computational resources motivates the topic of our research, the ramifications of this work could be quite broad. We demonstrate the power of combinatorial optimization techniques to handle such real-world problems. We also showcase the algorithm engineering process, consisting of the refinement of the mathematical model by an interaction of theoretical investigations and practical evaluations until an algorithm is found that matches all design goals. In addition, we develop a fully combinatorial linear-time algorithm for the TRANSSHIPMENT problem for the special graphs from our application, that may be of independent interest as it is also applicable to arbitrary graphs with bounded bandwidth and accordingly numbered nodes.

References

[1] F. Eisenbrand, A. Karrenbauer, M. Skutella, and C. Xu. Multiline addressing by network flow. In Y. Azar and T. Erlebach, eds., *Algorithms - ESA 2006, 14th Annual European Symposium*, Zürich, Switzerland, 2006, *LNCS 4168*, pp. 744–755. Springer.

- [2] F. Eisenbrand, A. Karrenbauer, M. Skutella, and C. Xu. Multiline addressing by network flow. *Algorithmica*, 53(4):583–596, 2009.
- [3] F. Eisenbrand, A. Karrenbauer, and C. Xu. Algorithms for longer oled lifetime. In C. Demetrescu, ed., *WEA 2007*, Rome, Italy, 2007, *LNCS 4525*, pp. 338–351. Springer.
- [4] A. Karrenbauer. *Engineering combinatorial optimization algorithms to improve the lifetime of OLED displays*. Phd thesis, Universität des Saarlandes, 2007.
- [5] C. Xu, A. Karrenbauer, K. M. Soh, and C. Codrea. Consecutive multiline addressing: A scheme for addressing pmoleds. *Journal of the Society for Information Display*, 16(2):211–219, 2008.
- [6] C. Xu, A. Karrenbauer, K. M. Soh, and J. Wahl. A New Addressing Scheme for PM OLED Display. In J. Morreale, ed., *SID 2007 International Symposium Digest of Technical Papers*, Long Beach, USA, 2007, vol. XXXVIII, pp. 97–100. Society for Information Display.

[7] C. Xu, J. Wahl, F. Eeisenbrand, A. Karrenbauer, K. M. Soh, and C. Hitzelberger. Verfahren zur ansteuerung von matrixanzeigen, 2005.

Recombination Networks

Investigators: Ernst Althaus and Rouven Naujoks

A fundamental task in computational biology is to reconstruct the ancestral relationship of a given set of species. In classic approaches this is done as a genealogy, that is, it is assumed that the course of evolution can be described by a tree. In order to build such a tree, specific features of the species are compared and the natural assumption is made that species with similar features are closely related. In modern phylogeny, such features are usually defined by DNA or protein sequences. One of the most widely used methods is to find a Steiner minimum tree in the Hamming metric, which is also referred to as *the maximum parsimony problem*. In [1], we have shown how to solve this NP-hard problem efficiently in practice.

Even though this restriction to tree-like structures has turned out to result in reliable reconstructions in many cases, in the presence of recombination events, that is, events that cause horizontal transfer of genetic data, such trees are not sufficient to reliably describe the ancestral relationship.

In [2], we propose a generalization of the maximum parsimony criterion for reconstructing phylogenetic networks under the assumption that such recombination events have occurred rarely. Furthermore, we describe an exact algorithm for one recombination event and show that in this case our method is not only able to reliably identify the recombined sequence but also to reconstruct the complete evolutionary history, which is in contrast to previous work like [3].

References

[1] E. Althaus and R. Naujoks. Computing steiner minimal trees in hamming metric. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'06*, Miami, USA, 2006, pp. 172–181. ACM / Siam.

[2] E. Althaus and R. Naujoks. Reconstructing phylogenetic networks with one recombination. In C. C. McGeoch, ed., *Experimental Algorithms, 7th International Workshop, WEA 2008*, Massachusetts, USA, 2008, *LNCS 5038*, pp. 275–288. Springer.

[3] J. Maydt and T. Lengauer. Recco: recombination analysis using cost optimization. *Bioinformatics*, 22(9):1064–1071, 2006.

Interval Constrained Coloring

Investigators: Ernst Althaus, Stefan Canzar, Khaled Elbassioni, Andreas Karrenbauer, and Julián Mestre

In [4] we introduced the *interval constrained coloring problem* as a mathematical abstraction of the problem of interpreting experimental data as obtained by monitoring, via mass spectrometry, the exchange of labile hydrogens for deuteriums (HDX) as a probe of protein surface accessibility. Despite several advantages HDX has over other laboratory-driven approaches to determine the tertiary structure of proteins, like X-ray crystal diffraction

and NMR, the HDX method involves a major difficulty: The output of the experiment only provides cumulative data for overlapping peptic fragments, the experiment does not deliver high-resolution data for single residues.

More formally, the interval constrained coloring problem can be formulated as follows: Let $\mathcal{I}$ be a set of intervals defined on the set $V = \{1, \ldots, n\}$, let $\mathcal{S} = \{1, \ldots, k\}$ be a set of color classes, and let $r : \mathcal{I} \times \mathcal{S} \mapsto \mathbb{Z}^+$ be a requirement function such that $\sum_{c \in \mathcal{S}} r(I, c) = |I|$ for all $I \in \mathcal{I}$. A coloring $\chi : V \mapsto \mathcal{S}$ is said to be *feasible* if for every $I \in \mathcal{I}$ we have

$$|\{i \in I \mid \chi(i) = c\}| = r(I, c) \text{ for all } c \in \mathcal{S}.$$

However, data collected in real experiments usually contain noise, which normally causes the instance to be infeasible. Therefore, we have captured the problem in [4] by an ILP, whose objective is to minimize the total sum of absolute deviations from the coloring requirements over all intervals. Based on the ILP formulation, we developed a *branch and bound* algorithm that was able to determine a coloring with minimum total error, for all real-world instances provided by our collaborator in biochemistry, in less than 0.1 second. Since biochemists are interested in *all* solutions with minimal error, we introduced a variant of the algorithm that enumerates a representative subset thereof.

We establish in [4] the polynomial-time solvability of the special case $k = 2$, i.e. when only two different exchange rates are distinguished, by the integrality of the linear programming relaxation polytope $\mathcal{P}$. We also propose a combinatorial polynomial-time algorithm for that case, which is based on a reformulation as a *minimum cost circulation* problem. We applied this algorithm as a subroutine to approximate solutions to instances with arbitrary but fixed number of colors and achieved an order of magnitude improvement in running time over the exact ILP approach. Furthermore, in [1] we used the combinatorial approach for the 2-color case to compute, in a Lagrangian fashion, a bound on the minimum total error, which is exploited in a branch-and-bound manner to determine all optimal colorings.

For $k \geq 3$, there are instances with fractional vertices, and the complexity status of the interval constrained coloring problem for three or more colors remains open. However, in [3] we show that deciding whether a feasible coloring exists is $\mathcal{NP}$-complete when k is part of the input. Nevertheless, we can check in polynomial time whether the linear relaxation $\mathcal{P}$ of the integer linear program is empty. Is that the case, there is clearly no feasible coloring. Otherwise we can find a feasible fractional solution. In [3], we show how to round this fractional solution to obtain a coloring that satisfies all the coloring requirements within a mere additive error of one. In light of our $\mathcal{NP}$-completeness result, this is essentially the best one can hope for.

To deal with the problem that the noise in the experimental data might force $\mathcal{P}$ to be empty, we also studied in [3] a variant of the problem that asks for a coloring that maximizes the number of requirements that are satisfied. More generally, given nonnegative weights $w(I)$ associated with intervals $I \in \mathcal{I}$, we seek a maximum weight subset of intervals for which a feasible coloring exists. We prove by a reduction from Max2Sat that this variant is $\mathcal{APX}$-hard for $k = 2$ and thus does not admit a polynomial-time approximation scheme (PTAS) unless $\mathcal{P} = \mathcal{NP}$. Therefore, we slightly relax the condition on when a requirement is satisfied and propose an approximation scheme which finds a coloring that "satisfies" the maximum number of coloring requirements and runs in quasi-polynomial time (QPTAS).

In the discussion so far, we took the coloring requirements for granted. Typically, the cumulative deuterium uptake data of peptic fragments as obtained from HDX experiments coupled with mass spectrometry are discretized in a preprocessing step into slow, medium and fast, i.e. $k = 3$. This can be achieved by fitting a rather simple model, using standard tools like least squares and maximum entropy methods (MEM) [5]. In [2], we went one step further and integrated the analysis of the cumulative deuterium uptake data and the assignment of exchange rates to single residues into a common mathematical optimization problem. Compared to the local decisions made by MEM to discretize the uptake data for single fragments, this global view on the experimental data eliminates a source of error.

More precisely, for an arbitrary but fixed number of exchange rates and given measurements of mass uptake at discrete time points, we wish to compute for each residue i an exchange rate k_i such that the expected deuterium uptake of a fragment at time t, assuming a certain underlying model of the exchange process, is close to the measured mass uptake at time t. Experiments on real-world instances show that optimal solutions to the resulting integer linear programming formulation provide exchange rates for single residues that lead to a close fit of the measured deuterium uptake. In particular, they show that our model is superior to the MEM analysis of single fragments. As already a relatively small number of exchange rates suffices to explain the data observed in the experiments well, our approach achieves reasonable running times.

References

[1] E. Althaus, S. Canzar, C. Ehrler, M. R. Emmett, A. Karrenbauer, A. G. Marshall, A. Meyer-Baese, J. Tipton, and H. Zhang. Computing h/d-exchange rates of single residues from data of peptic fragments. Submitted, 2009.

[2] E. Althaus, S. Canzar, C. Ehrler, M. R. Emmett, A. Karrenbauer, A. G. Marshall, A. Meyer-Baese, J. Tipton, and H. Zhang. Discrete fitting of hydrogen-deuterium-exchange-data of overlapping fragments. Submitted, 2009.

- [3] E. Althaus, S. Canzar, K. Elbassioni, A. Karrenbauer, and J. Mestre. Approximating the interval constrained coloring problem. In *11th Scandinavian Workshop on Algorithm Theory (SWAT)*, Gothenburg, Sweden, 2008, *LNCS 5124*, pp. 210–221. Springer.

- [4] E. Althaus, S. Canzar, M. Emmett, A. Karrenbauer, A. Marshall, A. Meyer-Bäse, and H. Zhang. Computing h/d-exchange speeds of single residues from data of peptic fragments. In *23rd Annual ACM Symposium on Applied Computing (SAC)*, Fortaleza, Brazil, 2008, pp. 1273–1277. ACM.

[5] Z. Zhang and D. L. Smith. Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. *Protein Sci.*, 2(4):522 – 531, 1993.

Fast and Accurate Bounds on Linear Programs

Investigators: Ernst Althaus and Daniel Dumitriu

State-of-the-art LP solvers rely on floating point arithmetic and hence can have wrong results. In [1], we give a method to certify the (in)feasibility of an LP or compute safe bounds on its value efficiently. We are considering LPs which are not too complicated numerically, and we want to find a certified answer with a bound close to the correct answer in most of the cases almost without overhead.

Previous approaches either take the basis computed by the LP solver and certify its feasibility and/or optimality with rational arithmetic (although if the basis is correctly found this computes very accurate bounds, in case the basis is infeasible, nothing can be said about the feasibility of the LP; it also takes much more time to solve a complicated LP with rational arithmetic), or compute bounds purely with floating point arithmetic (still for most of the LPs from Netlib no bound could be computed – for only 40% of them an upper bound could be computed).

To certify the feasibility of the linear system $\{A'x \leq b', A''x = b''\}$, we use a state-of-the-art LP solver to compute a feasible solution. Typically, we get a basis B such that $x^* = A_B^{-1} b_B$ has a unique solution, but not an exact solution vector, since we use floating point arithmetic. Then we use a solver for systems of linear equations that gives safe error bounds to obtain a vector of intervals $x^{\mathbb{I}} = (x^\ell, x^u) \in (\mathbb{R} \times \mathbb{R})^n$ with $A_B^{-1} b_B \in x^{\mathbb{I}}$. If all vectors in $x^{\mathbb{I}}$ satisfy strictly all remaining inequalities (checked using interval arithmetic), we are done. Unfortunately, many LPs are degenerated, i.e., some further inequalities are satisfied with equality. Therefore, we use the LP solver to find an optimal basis B for the LP $\max_{A'x + \mathbb{1}\delta \leq b', A''x = b''} \delta$, i.e., we try to find a feasible point of the linear system that is as far from tight at the inequalities as possible.

This modified approach typically works if δ is considerably larger than the error bound for the system of linear equations, but clearly fails, if the optimal value of the LP is 0. All inequalities with non-zero multipler in the dual LP should be satisfied with equality. Hence we transform all such inequalities to equalities and iterate. Notice that a wrong result of the floating point computations can lead to transforming too many inequalities to equalities, thus reducing the feasible region of the LP. Hence these fixings can lead to an infeasible LP although the original LP is feasible, but not the other way around. At least one of the new equalities is redundant and we want to safely remove it, i.e., we have to check that these equalities are indeed redundant with rational arithmetic. The same holds for redundant equalities that appear in the initial LP formulation. As typically the dual solution contains only a very small number of non-zero entries, the rational arithmetic is not too costly in this step.

Furthermore, if the LP solver returns a basis that contains an equality, this could be caused by a linear dependency among the linear equations (either in the original linear program or due to the transformation) and we have to certify whether this is the case before we remove these equalities. For the Netlib instances this is the case for 24 of the 94 LPs.

In a state-of-the-art LP solver, a highly tuned (in efficiency and numerical stability) algorithm for computing the LU decomposition is implemented. As the LP solver SoPlex allows us access to its LU decomposition, we want to use it in our computations. We do so by following approach of Highnam [2], where different estimates for $||L^{-1}||$ of an approximate LU decomposition are used to obtain rigorous error bounds.

In order to use this method, we have to detect and remove all redundant equalities. To remove an equality from our system, we have to make sure that it is indeed redundant. For this check, floating point arithmetic is not sufficient, since a wrong answer would make us remove an equality that is not redundant, and hence possibly enlarge the feasible region. Therefore we have to test redundancy with rational arithmetic. To show that $Ax = b$ contains a redundant equality one typically transforms $(A\ b)$ into row echelon form, an expensive operation, especially with rational arithmetic, therefore we have to use as few equations as

possible for which we compute the row echelon form.

When the objective value of the δ-modified LP is zero and an equality is redundant (i.e., the equality is in the basis), we have to transform the set of all equations into row echelon form. To improve the running time we can use a heuristic to select a subset of the equations so that this subset has as many redundant equalities as the set of all equalities; taking all equalities that are marked basic by the LP solver seems to be a good choice.

We were able to reduce the number of iterations by noticing that in many of them, a single inequality implies all bounds of the occurring variables. In this case, we simplify the LP (after double-checking with rational arithmetic), i.e., we set all occurring variables to the respective bounds and remove the inequality.

By computing the row echelon form of all equalities we can prove the feasibility of all but 3 LPs. If we compute the row echelon form of the equalities that are marked basic in the last LP, we can prove the feasibility of all but 9 LPs. In this case, the time spent using rational arithmetic is very small compared to the running time to solve the LPs in all but 3 LPs. Our simplification yields a big reduction in the number of calls to the LP solver. With our approach we are able to compute bounds for 90% of the Netlib instances, and the overhead in running time is negligible.

References

- [1] E. Althaus and D. Dumitriu. Fast and accurate bounds on linear programs. In *8th International Symposium on Experimental Algorithms (SEA 2009)*, Dortmund, Germany, 2009, Lecture Notes in Computer Science. Springer. Accepted for publication.

[2] N. J. Higham. A survey of condition number estimation for triangular matrices. *SIAM Review*, 29(4):575–596, 1987.

28.8.3 Combinatorial Algorithms

Here the aim is to develop more efficient algorithms that find optimal or approximate solutions for combinatorial optimization problems arising in different applications.

Cycle Bases in Graphs

Investigators: Tomasz Jurkiewics, Telikepalli Kavitha, Kurt Mehlhorn, and Dimitrious Michail in cooperation with Edoardo Amaldi, C. Iuliano (Milano), Christian Liebchen, Thorsten Ueckerdt(TU Berlin), Romeo Rizzi (Udine), and Katharina Zweig (Budapest)

Cycles in graphs play an important role in many applications, e.g., analysis of electrical networks, analysis of chemical and biological pathways, periodic scheduling, and graph drawing. From a mathematical point of view, cycles in graphs have a rich structure. Cycle bases are a compact description of the set of all cycles of a graph and cycle bases consisting of short cycles or, in weighted graphs, of small weight cycles are interesting mathematically and from an application viewpoint. In the applications above, sparse descriptions are to be preferred.

The study of cycle bases dates back to the early days of graph theory; [5] gave a characterization of planar graphs in terms of cycle bases. In the last decade, many new results on

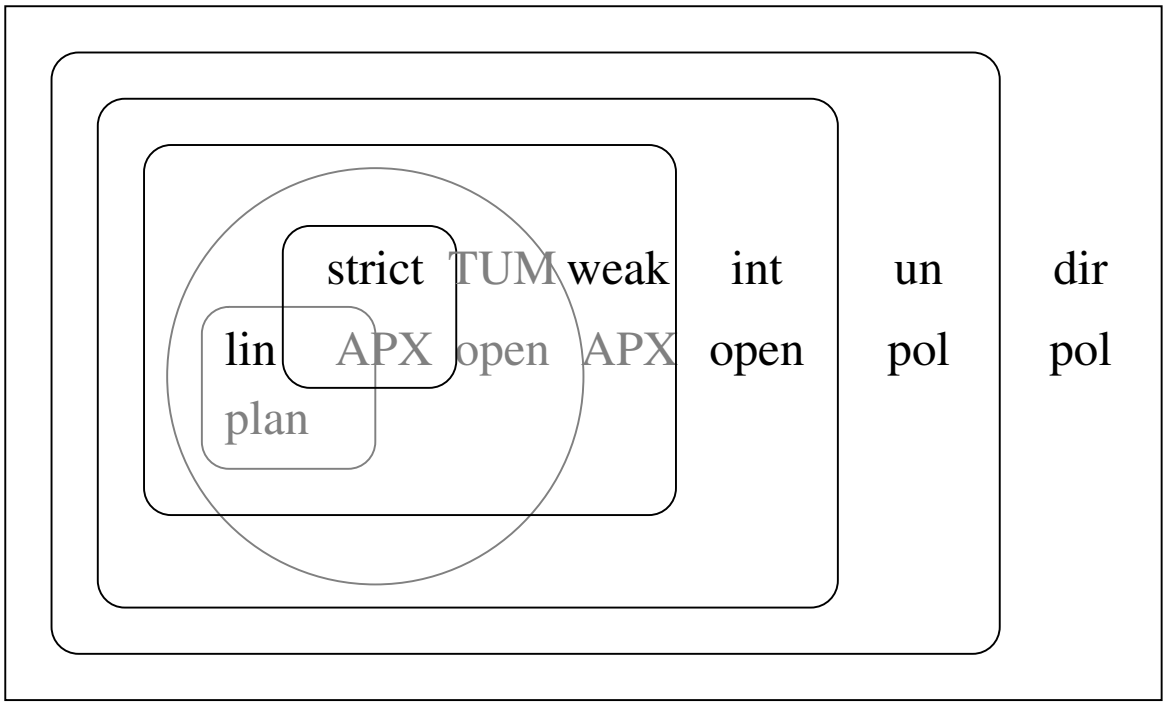


Figure 28.13: The inclusion diagram of cycle bases and the complexity status of their minimum weight cycle basis problems. The following abbreviations are used: dir = directed, un = undirected, int = integer, weak = weakly fundamental, TUM = totally unimodular, strict = strictly fundamental, plan = planar, pol = polynomial, APX = $\mathcal{APX}$-hard, lin = linear.

cycle bases appeared, most notably a classification of different kinds of cycle basis, structural results, a-priori bounds on the length and weight of minimum cycle bases, polynomial-time algorithms for constructing exact or approximate minimum cycle bases of some kinds and hardness results for other kinds of minimum cycle bases.

In [3], we survey these results and prove new ones. Figure 28.13 shows the landscape of cycle bases. We distinguish directed cycle bases, undirected cycle bases, integral cycle bases, weakly fundamental cycle basis, totally unimodular cycle bases, strictly fundamental cycle bases, and 2-bases. Each class can be characterized either combinatorially or algebraically. For example, undirected cycle bases are characterized by the fact that the determinant of their cycle matrix is odd and integral cycle basis are characterized by the fact that their determinant is ± 1. Figure 28.13 shows the inclusion map and the complexity status of the respective minimum cycle basis problem. Polynomial-time algorithms are known for directed, undirected and planar cycle bases. For weakly and strictly fundamental bases, the problem is $\mathcal{APX}$-hard and for integer and totally unimodular bases, the problem is open. Every graph of n nodes and m edges has a weakly fundamental cycle basis of length $O(m \log m / \log(m/n))$. There are graphs that have no shorter basis. Different applications need different kinds of

bases, e.g., the analysis of electrical circuits can do with any kind of cycle basis, periodic scheduling requires integral cycle bases, and graph drawing needs strictly fundamental bases.

Until recently, the best algorithm for undirected bases [7] had running time $O(\frac{m^2n}{\log n} + mn^2)$. For directed bases [4, 6], the best deterministic algorithm has running time $O(m^3n)$ and the best Monte Carlo algorithm had running time $O(m^2n)$. The algorithms work only for non-negative edge weights.

In [3], we give the first algorithms that work for conservative weight functions. A weight function for an undirected graph is conservative iff every circuit has nonnegative weight. For undirected bases, we obtain running time $O(n^3 \log n + \frac{m^2n}{\log n} + mn^2)$, and for directed bases, we obtain deterministic time $O(m^3n)$ and Monte Carlo time $O(n^3 \log n + m^2n)$.

In very recent work [1], we obtain $O(m^\omega)$ Monte Carlo algorithms for minimum undirected and directed bases and non-negative edge weights. For sparse graphs with $m = O(n)$, this is an improvement by the factor $n^{3-\omega}$. More importantly, we give the *first* algorithm for sparse graphs that is faster than $O(n^3)$. We should mention that the previously fastest algorithms already used fast matrix multiplication.

Hartvigsen and Mardon [2] have shown that minimum undirected cycle bases in planar graphs can be computed in time $O(n^2 \log n)$. We [1] improve the running time to $O(n^2)$. As a consequence, the all-pairs minimum cut problem in planar graphs can also be solved in time $O(n^2)$. As a structural result, we show that every planar graph has a minimum directed cycle basis that is weakly fundamental, totally unimodular, and integral.

References

- [1] E. Amaldi, C. Iuliano, T. Jurkiewics, K. Mehlhorn, and R. Rizzi. Improved algorithms for cycle basis. 2009.

[2] D. Hartvigsen and R. Mardon. The all-pairs min cut problem and the minimum cycle basis problem on planar graphs. *SIAM J. Discrete Math.*, 7(3):403–418, 1994.

- [3] T. Kavitha, C. Liebchen, K. Mehlhorn, D. Michail, R. Rizzi, T. Ueckerdt, and K. Zweig. Cycle bases in graphs: Characterization, algorithms, complexity, and applications. 2009.
- [4] T. Kavitha and K. Mehlhorn. Algorithms to compute minimum cycle basis in directed graphs. *Theory of Computing Systems*, 40(4):485–505, 2007.

[5] S. MacLane. A combinatorial condition for planar graphs. *Fundementa Mathematica*, 28:22–32, 1937.

- [6] K. Mehlhorn and D. Michail. Implementing minimum cycle basis algorithms. *Journal of Experimental Algorithmics*, 11:1–14, 2007.
- [7] K. Mehlhorn and D. Michail. Minimum cycle bases: Faster and simpler. *ACM Transactions on Algorithms*, 2009. accepted for publication.

Assigning Papers to Referees

Investigators: Kurt Mehlhorn and Julián Mestre in cooperation with Naveen Garg, Amit Kumar, and Kavitha Telikepalli

Refereed conferences require every submission to be reviewed by members of a program committee (PC) in charge of selecting the conference program. There are many software

packages available to manage the review process. Typically, in a bidding phase PC members express their personal preferences by ranking the submissions. This information is used by the system to compute an assignment of the papers to referees (PC members).

In [1], we studied the problem of *assigning papers to referees*. We proposed to optimize a number of criteria that aim at achieving fairness among referees/papers. Some of these variants can be solved optimally in polynomial time, while others are NP-hard, in which case we designed approximation algorithms. Experimental results strongly suggest that the assignments computed by our algorithms are considerably better than those computed by the popular conference management software EasyChair [2].

In the next phase of this project, we plan to perform a thorough experimental evaluation of our algorithms and eventually incorporate them into EasyChair.

References

- [1] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre. Assigning papers to referees. 2008.

[2] A. Voronkov. Easychair. `http://www.easychair.org`.

Popular Mixed Matchings

Investigators: Julián Mestre in cooperation with Kavitha Telikepalli and Meghana Nasre

Consider the problem of matching applicants to jobs under one-sided preferences; that is, each applicant ranks a non-empty subset of jobs under an order of preference, possibly involving ties. A matching M is said to be *more popular* than T if the applicants that prefer M to T outnumber those that prefer T to M. A matching is said to be *popular* if there is no matching more popular than it. Equivalently, a matching M is popular if $\phi(M,T) \geq \phi(T,M)$ for all matchings T, where $\phi(X,Y)$ is the number of applicants that prefer X to Y.

Previously studied solution concepts based on the popularity criterion are either not guaranteed to exist for every instance (e.g., popular matchings [1]) or are NP-hard to compute (e.g., least unpopular matchings [3]). In [2] we addressed this issue by considering mixed matchings. A *mixed matching* is simply a probability distributions over matchings in the input graph. The function ϕ that compares two matchings generalizes in a natural manner to mixed matchings by taking expectation. A mixed matching P is popular if $\phi(P,Q) \geq \phi(Q,P)$ for all mixed matchings Q.

We showed that popular mixed matchings *always* exist and we design polynomial-time algorithms for finding them. Then we studied their efficiency and gave tight bounds on the price of anarchy and price of stability of the popular matching problem.

References

[1] D. J. Abraham, R. W. Irving, T. Kavitha, and K. Mehlhorn. Popular matchings. *SIAM Journal on Computing*, 37(4):1030–1045, 2007.

- [2] T. Kavitha, J. Mestre, and M. Nasre. Popular mixed matchings. In *36th International Colloquium on Automata, Languages and Programming*, Rhodes, Greece, 2009. Springer.

[3] R. M. McCutchen. The least-unpopularity-factor and least-unpopularity-margin criteria for matching problems with one-sided preferences. In *Proceedings of the 15th Latin American Symposium on Theoretical Informatics*, 2008, pp. 593–604.

On Short Paths Interdiction Problems: Total and Node-wise Limited Interdiction

Investigators: Khaled Elbassioni in cooperation with Endre Boros, Konrad Borys, Vladimir Gurvich, Leonid Khachiyan, Gabor Rudolf, and Jihui Zhao

In [1], we consider the following problem. Given a directed graph $G = (V, A)$ with a non-negative weight (length) function on its arcs $w : A \to \mathbb{R}_+$ and two terminals $s, t \in V$, the goal is to destroy all short directed paths from s to t in G by eliminating some arcs of A. This is known as the *short paths interdiction problem*. We consider several versions of it, and in each case analyze two subcases: *total limited interdiction*, when a fixed number k of arcs can be removed, and *node-wise limited interdiction*, when for each node $v \in V$ a fixed number $k(v)$ of out-going arcs can be removed. Our results indicate that the latter subcase is always easier than the former one. In particular, we show that the short paths node-wise interdiction problem can be efficiently solved by an extension of Dijkstra's algorithm. In contrast, the short paths total interdiction problem is known to be NP-hard. We strengthen this hardness result by deriving the following inapproximability bounds: Given k, it is NP-hard to approximate within a factor $c < 2$ the maximum $s - t$ distance $d(s, t)$ obtainable by removing (at most) k arcs from G. Furthermore, given d, it is NP-hard to approximate within a factor $c < 10\sqrt{5} - 21 \approx 1.36$ the minimum number of arcs which has to be removed to guarantee $d(s, t) \geq d$. Finally, we also show that the same inapproximability bounds hold for undirected graphs and/or node elimination.

References

- [1] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, V. Gurvich, G. Rudolf, and J. Zhao. On short paths interdiction problems: Total and node-wise limited interdiction. *Theory Comput. Syst.*, 43(2):204–233, 2008.

On the Design of Efficient Interruptible Algorithms

Investigators: Spyros Angelopoulos in cooperation with Alejandro López-Ortiz and Angele Hamel

Anytime algorithms are algorithms whose quality of output improves gradually as the available computation time increases. These algorithms have been proven extremely useful in AI applications, especially in situations where the available computation time is not known in advance. Such applications include, for instance, medical diagnostic systems, motion planning under kinetic constraints and game-playing systems (see, e.g. the survey of Zilberstein [3]).

One can distinguish between two types of anytime algorithms. On one hand, *interruptible algorithms* can be interrupted at any point throughout their execution, at which point they must report their current solution. On the other hand, the class of *contract algorithms* consists of algorithms for which the allowable execution time is part of their input. If a contract algorithm is queried before its alloted execution time elapses, it might not yield any useful

results. A central topic in the design of real-time systems is converting any contract algorithm to its interruptible version. This can be accomplished, in black-box fashion, by scheduling executions of contract algorithms on either a single or multiple processors. At interruption time, the result of the "best" completed execution of a contract algorithm is returned as the solution. The problem has been studied intensively within the AI community, and in earlier work with A. López-Ortiz and A. Hamel [2] we presented an optimal scheduling strategy for the general problem.

In more recent work [1], we considered the setting in which the interruption time t is not a hard deadline for reporting a solution, but instead a "grace period" $w(t)$ is granted, within which the algorithm must report a solution. Hence, in this setting, interruptions are treated as "soft" deadlines, which offers some flexibility in an otherwise stringent real-time system. We proposed measures that capture the efficiency of schedules in the presence of soft deadlines, and provided optimal schedules for each such measure.

References

- [1] S. Angelopoulos, A. Lopez-Ortiz, and A. Hamel. Optimal scheduling of contract algorithms with soft deadlines. In *23rd National Conference on Artificial Intelligence (AAAI 2008)*, Chicago, 2008, pp. 868–873. AAAI press.

[2] A. López-Ortiz, S. Angelopoulos, and A. Hamel. Optimal scheduling of contract algorithms for anytime problems. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006.

[3] S. Zilberstein. Using anytime algorithms in intelligent systems. *AI Magazine*, 17(3):73–83, 1996.

28.8.4 Computational Issues Related to Transversal Hypergraphs and Polyhedra

Many practical applications in different areas require the listing of combinatorial objects or structures satisfying certain (monotone) properties. We consider two fundamental problems in this area, whose complexities remain very challenging: *generating hypergraphs transversals* and *vertex enumeration for polyhedra*, and other related problems.

Hypergraph Transversals

Investigators: Khaled Elbassioni and Imran Rauf in cooperation with Endre Boros, Matthias Hagen, and Kazuhisa Makino

Let V be a finite set of vertices, and $\mathcal{H} \subseteq 2^V$ be a hypergraph (a family of subsets) on V. A *transversal* or *hitting set* of $\mathcal{H}$ is a subset of vertices that has non-empty intersection with every subset in the family $\mathcal{H}$. The *transversal hypergraph* of $\mathcal{H}$, denoted as $\mathcal{H}^d$, is then defined as the family of all inclusion-wise minimal transversals of $\mathcal{H}$. It can be shown that for Sperner hypergraphs (i.e., no subset in the family contains another) $\mathcal{H}^{dd} = \mathcal{H}$ and so $\mathcal{H}^d$ is also called the *dual* of $\mathcal{H}$.

Given a hypergraph $\mathcal{H}$, the *hypergraph transversal* or *dualization* problem asks for generating the transversal hypergraph $\mathcal{H}^d$. It can be easily seen that the output size can be exponential in the input size, so it is more reasonable to consider *output-sensitive* polynomial-time algorithms

for this problem. While the current best known bound for the problem is quasi-polynomial ($O(n^{o(\log n)})$) [7], where n is the combined input-output size, many interesting classes of hypergraphs are known to be dualizable in polynomial time.

For a positive integer r, we call hypergraph $\mathcal{H}$ *r-exact*, if any minimal transversal of $\mathcal{H}$ intersects any hyperedge of $\mathcal{H}$ in at most r vertices. This class includes several interesting examples from geometry, e.g., circular-arc hypergraphs ($r = 2$), hypergraphs defined by sets of axis-parallel lines stabbing a given set of α-fat objects ($r = 4\alpha$, see Figure 28.14), and hypergraphs defined by sets of points contained in translates of a given cone in the plane ($r = 2$). For constant r, we give in [6] a polynomial-time algorithm for the duality testing problem of a pair of r-exact hypergraphs. This result implies that minimal hitting sets for the above geometric hypergraphs can be generated in output-sensitive polynomial time.

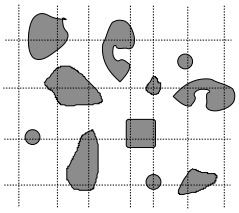


Figure 28.14: Stabbing α-fat objects in Euclidean plane

In [4], we give fixed-parameter algorithms for generating transversal hypergraphs and related problems. In the first part of this work, we consider the number of edges of the hypergraphs, the maximum degree of a vertex, and maximum vertex complementary degree as our parameters. In the second part, we use an *a priori*-based approach to obtain fixed-parameter tractability results for generating all maximal independent sets of a hypergraph, all minimal transversals of a hypergraph, and all maximal frequent sets where parameters bound the intersections or unions of hyperedges.

Equivalently, the hypergraph dualization problem can be stated as that of finding the *irredundant disjunctive normal* form (DNF) ϕ representation of a *monotone* Boolean function $f : \{0,1\}^n \to \{0,1\}$, given its *irredundant conjunctive normal* form (CNF) ψ. A very simple method, sometimes called *Berge multiplication*, works by multiplying out the clauses of ϕ from left to right in some order, simplifying whenever possible using *the absorption law*. In [1], we show that for any monotone CNF ϕ, Berge multiplication can be done in subexponential time, and for many interesting subclasses of monotone CNF's such as CNF's with bounded size, bounded degree, bounded intersection, bounded conformality, and read-once formula, it can be done in polynomial or quasi-polynomial time.

In [5], we consider the following related problem. The *readability* of a monotone Boolean function f is defined as the minimum integer k such that there exists an $\wedge - \vee$-formula equivalent to f in which each variable appears at most k times. One of the earliest results in this direction is the characterization of read-once functions by Gurvich [9, 10]. Based on these criteria, Golumbic [8] presents an algorithm to recognize Conjunctive or Disjunctive Normal Forms (CNF/DNF) of read-once functions. A natural next step is to ask whether Gurvich's characterization can be directly generalized for read-twice functions but this is shown to be not true in our work [5]. If the input representation is not in CNF/DNF form, but rather a general monotone Boolean formula, then we show that it is NP-hard to decide if a given monotone formula represents a read-once function. It follows also from our reduction that it is NP-hard to approximate the readability of a given monotone Boolean function $f : \{0,1\}^n \to \{0,1\}$ within a factor of $O(n)$.

On the positive side, we give tight sublinear upper bounds on the readability of a monotone Boolean function given in CNF (or DNF) form, parameterized by the number of terms in

the CNF and the maximum size in each term, or more generally the maximum number of variables in the intersection of any constant number of terms. When the variables of the DNF can be ordered so that each term consists of a set of consecutive variables, we give much tighter polylogarithmic bounds on the readability.

In [3], we consider a generalization of the basic hypergraph transversal problem, which has several interesting applications in different areas, e.g. data mining. Let $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ be the product of n partially ordered sets. Given a subset $\mathcal{A} \subseteq \mathcal{P}$, we consider problem $DUAL(\mathcal{P}, \mathcal{A}, \mathcal{B})$ of extending a given partial list $\mathcal{B}$ of maximal independent elements of $\mathcal{A}$ in $\mathcal{P}$. In [3], we give quasi-polynomial-time algorithms for solving problem $DUAL(\mathcal{P}, \mathcal{A}, \mathcal{B})$ when each poset $\mathcal{P}_i$ belongs to one of the following classes: (i) semi-lattices of bounded width, (ii) forests, that is, posets with acyclic underlying graphs, with either bounded in-degrees or out-degrees, or (iii) lattices defined by a set of real closed intervals.

As an application, we consider a system $A \circ x \geq b$, where $A \in \mathbb{R}_+^{m \times n}$ is a non-negative matrix and $b \in \mathbb{R}_+^m$ is a non-negative vector over the n-dimensional variable $l \leq x \leq u$, where $l, u \in \mathbb{R}_+^n$ are lower and upper bounds respectively, and $\circ$ is either a max-min or a max-product composition. In the special case, when every component of A, b and x is in $[0, 1]$, and the inequality is replaced by an equality, the problem is known in the literature as *Fuzzy relations equation constraints*. By looking at this problem as a monotone generation problem in the products of total orders, we show in [2] that the set of minimal solutions of such systems can be computed in incremental quasi-polynomial time.

References

- [1] E. Boros, K. Elbassioni, and K. Makino. On berge multiplication for monotone boolean dualization. In *Automata, Languages and Programming, 35th International Colloquium, ICALP 2008*, Reykjavik, Iceland, 2008, *LNCS 5125*, pp. 48–59. Springer.
- [2] K. Elbassioni. A note on systems with max-min and max-product constraints. *Fuzzy Sets and Systems*, 159(17):2272–2277, 2008.
- [3] K. Elbassioni. Algorithms for dualization over products of partially ordered sets. *SIAM J. Discrete Math*, 23(1):487–510, 2009.
- [4] K. Elbassioni, M. Hagen, and I. Rauf. Some fixed-parameter tractable classes of hypergraph duality and related problems. In *Parameterized and Exact Computation, Third International Workshop, IWPEC 2008*, Victoria, Canada, 2008, *LNCS 5018*, pp. 91–102. Springer.
- [5] K. Elbassioni, K. Makino, and I. Rauf. On the readability of monotone boolean formulae. (To appear in COCOON 2009), 2009.
- [6] K. Elbassioni and I. Rauf. Polynomial-time dualization of r-exact hypergraphs with applications in geometry. (submitted to a journal), 2009.
- [7] M. Fredman and L. Khachiyan. On the complexity of dualization of monotone disjunctive normal forms. *J. Algorithms*, 21(3):618–628, 1996.
- [8] M. C. Golumbic, A. Mintz, and U. Rotics. Factoring and recognition of read-once functions using cographs and normality and the readability of functions associated with partial k-trees. *Discrete Applied Mathematics*, 154(10):1465–1477, 2006.
- [9] V. Gurvich. On repetition-free boolean functions. *Uspekhi Mat. Nauk. (Russian Math. Surveys)*, 32:183–184, 1977. (in Russian).

[10] V. Gurvich. Criteria for repetition-freeness of functions in the algebra of logic. *Soviet Math. Dokl.*, 43(3):721–726, 1991.

Vertex Enumeration of Polyhedra and Related Problems

Investigators: Khaled Elbassioni in cooperation with Endre Boros, Vladimir Gurvich, and Hans Raj Tiwary

A polytope $P \subseteq \mathbb{R}^n$ is the convex hull of a finite number of points in $\mathbb{R}^n$. A very basic result in the theory of polytopes states that every polytope can also be represented as the intersection of a finite number of halfspaces. These representations are commonly referred to as the $\mathcal{V}$- and $\mathcal{H}$-representations, respectively, and are unique assuming that none of the vertices or halfspaces are redundant. Converting from the $\mathcal{V}$-representation to the $\mathcal{H}$-representation is the well-known *convex hull* problem (CH), and from the $\mathcal{H}$-representation to the $\mathcal{V}$-representation is the well-known *vertex enumeration* problem (VE). These two problems are polynomially equivalent via linear programming, and are among the most important longstanding open questions in the theory of polytopes.

Motivated by the goal to show that the vertex enumeration of polytopes is a problem that somehow behaves similarly to the hypergraph dulalization problem mentioned above, we study in [8] the complexity of determining whether a polytope given by its vertices or facets is combinatorially isomorphic to its polar dual. We prove that this problem is Graph Isomorphism hard, and that it is Graph Isomorphism complete if and only if vertex enumeration is Graph Isomorphism easy. To the best of our knowledge, this is the first problem that is not equivalent to vertex enumeration and whose complexity status has a non-trivial impact on the complexity of vertex enumeration, irrespective of whether checking Self-duality turns out to be strictly harder than Graph Isomorphism or equivalent to Graph Isomorphism. The constructions employed in the proof yield a class of self-dual polytopes that are interesting on their own. In particular, this class of self-dual polytopes has the property that the facet-vertex incident matrix of the polytope is transposable if and only if the matrix is symmetrizable as well. As a consequence of this construction, we also prove that checking self-duality of a polytope, given by its facet-vertex incidence matrix, is Graph Isomorphism complete, thereby answering a question of Kaibel and Schwartz [6].

Given a graph $G = (V, E)$ and a weight function on the edges $w : E \mapsto \mathbb{R}$, we consider in [2] the polyhedron $P(G, w)$ of negative-weight flows on G, and get a complete characterization of the vertices and extreme directions of $P(G, w)$. Based on this characterization, and using a construction developed in [7], we show that, unless $P = NP$, there is no output polynomial-time algorithm to generate all the vertices of a 0/1-polyhedron. This strengthens the NP-hardness result of [7] for non 0/1-polyhedra, and comes in contrast with the polynomiality of vertex enumeration for 0/1-polytopes [3]. As further applications, we show that it is NP-hard to check if a given integral polyhedron is 0/1, or if a given polyhedron is half-integral. We also show that it is NP-hard to approximate the maximum support of a vertex a polyhedron in $\mathbb{R}^n$ within a factor of $\Omega(1/n)$. Finally, with further developments in [5], we show that it is NP-hard to approximate the vertex centroid (that is the average of the vertices) of a given $\mathcal{H}$-polyhedron within any non-trivial distance.

In [4] we consider the following related problem. Given a set of polyhedral cones $\mathcal{C}_1, \cdots, \mathcal{C}_k \subset \mathbb{R}^d$, and a convex set D, does the union of these cones cover the set D? We consider the

computational complexity of this problem for various cases such as whether the cones are defined by extreme rays or facets, and whether D is the entire $\mathbb{R}^d$ or an affine subspace $\mathbb{R}^t$. As a consequence, we show that the problem of checking if the union of a given set of convex polytopes is convex is coNP-complete, thus answering a question of Bemporad et al. [1].

References

[1] A. Bemporad, K. Fukuda, and F. D. Torrisi. Convexity recognition of the union of polyhedra. *Comput. Geom.*, 18(3):141–154, 2001.

- [2] E. Boros, K. Elbassioni, V. Gurvich, and H. R. Tiwary. Characterization of the vertices and extreme directions of the negative cycle polyhedron and hardness of generating vertices of 0/1-polyhedra. *CoRR*, abs/0801.3790, 2008.

[3] M. R. Bussieck and M. E. Lübbecke. The vertex set of a 0/1 polytope is strongly $\mathcal{P}$-enumerable. *Computational Geometry: Theory and Applications*, 11(2):103–109, 1998.

- [4] K. Elbassioni and H. R. Tiwary. On a cone covering problem. In *Proceedings of the 20th Annual Canadian Conference on Computational Geometry*, Montreal, Canada, 2008, pp. 171–174. CCCG.

- [5] K. Elbassioni and H. R. Tiwary. On computing the vertex centroid of a polyhedron. *CoRR*, abs/0806.3456, 2008.

[6] V. Kaibel and A. Schwartz. On the complexity of polytope isomorphism problems. *Graphs and Combinatorics*, 19(2):215–230, 2003.

- [7] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, and V. Gurvich. Generating all vertices of a polyhedron is hard. *Discrete & Computational Geometry*, 39(1-3):174–190, 2008.

- [8] H. R. Tiwary and K. Elbassioni. On the complexity of checking self-duality of polytopes and its relations to vertex enumeration and graph isomorphism. In *Symposium on Computational Geometry 2008*, College Park, MD, USA,, 2008, pp. 192–198. ACM.

A Complete Characterization of Nash-Solvability of Bimatrix Games in Terms of the Exclusion of Certain 2×2 Subgames

Investigators: Khaled Elbassioni in cooperation with Endre Boros, Vladimir Gurvich, Kazuhisa Makino, and Vladimir Oudalov

This is yet another interesting application of the hypergraph transversal problem, or its equivalent, *joint generation*.

In 1964 Shapley observed that a matrix has a saddle point whenever every 2×2 submatrix of it has one. In contrast, a bimatrix game may have no Nash equilibrium (NE) even when every 2×2 subgame of it has one. Nevertheless, Shapley's claim can be generalized for bimatrix games in many ways as follows. We partition all 2×2 bimatrix games into fifteen classes $C = \{c_1, \ldots, c_{15}\}$ depending on the preference pre-orders of the two players. A subset $t \subseteq C$ is called a NE-theorem if a bimatrix game has a NE whenever it contains no subgame from t. In [1], we suggest a general method for getting all minimal (that is, strongest) NE-theorems based on the procedure of joint generation of transversal hypergraphs given by a special oracle. By this method we obtain all (six) minimal NE-theorems.

References

- [1] E. Boros, K. Elbassioni, V. Gurvich, K. Makino, and V. Oudalov. A complete characterization of nash-solvability of bimatrix games in terms of the exclusion of certain 2x2 subgames. In *Computer Science - Theory and Applications, Third International Computer Science Symposium in Russia, CSR 2008*, Moscow, Russia, 2008, vol. 5010, pp. 99–109. Springer.

28.9 Algorithmic Game Theory and Online Algorithms

Coordinator: Rob van Stee

Many classical combinatorial optimization problems take on a new flavor when considered on the Internet. Classically, we usually assume that we are given some input (in a block, so that we then have all the data that we need), and calculate (fix) some output or assignment in a centrally controlled manner, without having to deal with outside agents. By contrast, the Internet is by its nature distributed, asynchronous, uncoordinated, and divided into many sites or areas that are controlled by agents that are principally interested in their own welfare. Therefore, if we want to carry out projects on the Internet that involve several or many such agents, we need to ensure that cooperation is in the best interest of these agents, since we cannot control their behavior deterministically. Moreover, we cannot in general be sure that the data that these agents provide us with is accurate, especially if they could somehow benefit from giving us false information.

Many new and interesting problems arise in this setting, and our group investigates such problems from two directions. The first is that of *algorithmic mechanism design*, which focuses on getting selfish agents to reveal their true (private) data in order to optimize performance. This requires using an auction mechanism or (more often) designing suitable (monotone) approximation algorithms for the problem at hand. Our main result in the reporting period is a lower bound of $1 + \sqrt{2}$ on the approximation ratio of any monotone (and hence truthful) mechanism for scheduling tasks on n unrelated machines. It is a major open problem to close the gap between the lower and the upper bound. See Section 28.9.1 for details. In the context of the Internet, where users appear and disappear continuously, it is also natural and important to consider *online* algorithms for such problems. In particular, we are currently working on *prompt* mechanisms for auctions, that are required to set the price when the item is sold instead of waiting until the end.

We are also interested in directly examining the effects of selfishness in combinatorial optimization. That is, we consider Nash equilibria, which are stable states in which no agent has a motive to deviate from their strategy, and compare them to the best possible solution that could be reached in a centralized setting, where there is a single controller and no freedom for the agents. Our main result here is a proof that Nash equilibria are not necessarily unique in routing games with players who control large amounts of flow and who can route their flow fractionally (Section 28.9.3).

Understanding the worst-case distance of a Nash equilibrium from the social optimum in simple situations is also a prerequisite for making progress in the area of (algorithmic) mechanism design. This worst-case distance is known as the coordination ratio or the *price of anarchy* (POA). Taking a more positive point of view, the *price of stability* is the ratio of *best* equilibrium and the best optimal design. The best Nash equilibrium is the best solution

that can be proposed from which no user will 'defect'. One of the problems we are currently interested in is determining the price of stability for undirected network design with fair cost sharing, in particular, whether this price grows with the number of players or not.

The second major research topic of our group is *online algorithms*. In an online problem, the input arrives incrementally and irrevocable decisions need to be made before the next part of the input arrives. This models many real-world problems where it is infeasible to wait until the entire input has arrived before making any decisions. We compare the results that can be obtained under such conditions to the best possible solutions when the entire input is given in advance. Thus the focus here is on the cost of a lack of information, in contrast to the area of approximation algorithms, where we focus on the effect of having limited (i.e., polynomial) computation time. We have two main results. First, we give a perhaps ultimate analysis of the well-known algorithm LRU for paging. We applied a powerful, yet simple and intuitive form of analysis termed *bijective analysis* for the paging and list update problems. Our main result is a theoretical justification of the superiority of the Least-Recently-Used strategy, which is precisely what is observed in practice. Second, we showed an $\omega(1)$ lower bound on the competitive ratio of any deterministic online algorithm for minimizing the total weighted flow time on a single machine with preemptions, disproving a widely held belief. For more details see Section 28.9.4.

28.9.1 Algorithmic Mechanism Design

Minimizing the Makespan on Unrelated Machines

Investigator: Giorgos Christodoulou

Scheduling on unrelated machines is a classical NP-hard problem. The problem is one of the most fundamental scheduling problems [6]. There are n machines and m tasks and each task may have different execution times on the machines. Let t_{ij} be execution time of task j on machine i. The objective is to schedule the tasks on the machines to minimize the makespan. Lenstra, Shmoys and Tardos [6] gave a 2-approximation polynomial time algorithm, while they also proved that the problem cannot be approximated (in polynomial time) within a factor less than 3/2.

The mechanism design version of the problem originates in the seminal work of Nisan and Ronen [8]. In the mechanism design setting, each machine i knows its own times (the t_{ij}'s), but the algorithm does not know them. The algorithm first asks the machines to declare their times t_{ij} and then proceeds to allocate the tasks according to a policy known to machines in advance. The machines are selfish players who are lazy and do not want to execute the tasks, so they may lie. To deal with this problem, the mechanism pays the machines according to their declarations. Thus the mechanism design problem consists of two algorithms: an allocation algorithm and a payment algorithm. They take as input the declaration of times by the machines and produce an allocation and a set of payments, one for each machine. Nisan and Ronen [8] gave a n-approximation deterministic truthful mechanism and a lower bound of 2. They also conjectured the actual bound to be n. In [4], we improve the lower bound to $1 + \sqrt{2}$ for 3 or more machines. For a survey of results concerning the problem see [1].

In [3] we completely characterize the decisive truthful mechanisms for this problem for two players when the domain contains both positive and negative values. We show that the

class of truthful mechanisms is very limited: A decisive truthful mechanism partitions the tasks into groups so that the tasks in each group are allocated independently of the other groups. Tasks in a group of size at least two are allocated by an affine minimizer and tasks in singleton groups by a task-independent mechanism.

A direct consequence of this approach is that the approximation ratio of mechanisms for two players, for the objective of minimizing the makepan, is 2, even for two tasks. In fact, it follows that for two players, VCG is the unique algorithm with optimal approximation 2.

In [2], we consider fractional scheduling. We give a $2-1/n$ lower bound on the approximation ratio, that can be achieved by any truthful mechanism. This result shows that even in the case of such a problem, for which the offline version can be exactly solved in polynomial time, its mechanism design analog may turn out to be impossible to approximate, even by non-polynomial mechanisms. Note that giving a lower bound for fractional mechanisms is another way to obtain lower bounds for randomized mechanisms for the integral case. Consequently, our $2 - 1/n$ lower bound extends the lower bounds in [7] to the class of fractional mechanisms.

In the positive direction, we give a truthful mechanism with approximation ratio 3/2 for 2 machines, which matches our lower bound. The generalization of our mechanism for n machines gives us an approximation ratio of $1 + \frac{n-1}{2}$. We show that this is optimal among *task-independent* mechanisms, where the decision for the assignment of a task depends only on the processing times that concern the particular task. Considering task-independence is motivated by the fact that all known 'reasonable' deterministic and randomized mechanisms for this problem are task-independent.

Maximizing the Minimum Load on Related Machines

Investigator: Rob van Stee

We consider the problem of maximizing the minimum load for machines that are controlled by selfish agents, who are only interested in maximizing their own profit [5]. Unlike the classical load balancing problem, this problem has not been considered for selfish agents until now. We consider the version where the machines are related. Hence, each agent has a single private value, which corresponds to the speed of its machine.

For a constant number of machines, m, we show a monotone (and hence truthful) polynomial time approximation scheme (PTAS) with running time that is linear in the number of jobs. It uses a new technique for reducing the number of jobs while remaining close to the optimal solution. We also present an FPTAS for the classical problem, i.e., where no selfish agents are involved (the previous best result for this case was a PTAS) and use this to give a monotone FPTAS.

Additionally, we give a monotone approximation algorithm with approximation ratio $\min(m, (2+\varepsilon)s_1/s_m)$ where $\varepsilon > 0$ can be chosen arbitrarily small and s_i is the (real) speed of machine i. Finally we give improved results for two machines.

References

- [1] G. Christodoulou and E. Koutsoupias. Mechanism design for scheduling. *Bulletin of European Association for Theoretical Computer Science*, 2009.

- [2] G. Christodoulou, E. Koutsoupias, and A. Kovács. Mechanism design for fractional scheduling on unrelated machines. *ACM Transactions on Algorithms*, 2009.
- [3] G. Christodoulou, E. Koutsoupias, and A. Vidali. A characterization of 2-player mechanisms for scheduling. In D. Halperin and K. Mehlhorn, eds., *Algorithms - ESA 2008, 16th Annual European Symposium, Karlsruhe, Germany, September 15-17, 2008. Proceedings.*, Karlsruhe, Germany, 2008, pp. 297–307. Springer.
- [4] G. Christodoulou, E. Koutsoupias, and A. Vidali. A lower bound for scheduling mechanisms. *Algorithmica*, 2009.
- [5] L. Epstein and R. van Stee. Maximizing the minimum load for selfish agents. In E. S. Laber, C. Bornstein, L. T. Noguiera, and L. Faria, eds., *LATIN 2008: Theoretical Informatics, 8th Latin American Symposium*, Búzios, Brazil, 2008, *LNCS 4957*, pp. 264–275. Springer.

[6] J. Lenstra, D. Shmoys, and É. Tardos. Approximation algorithms for scheduling unrelated parallel machines. *Mathematical Programming*, 46(1):259–271, 1990.

[7] A. Mu'alem and M. Schapira. Setting lower bounds on truthfulness. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2007, pp. 1143–1152.

[8] N. Nisan and A. Ronen. Algorithmic mechanism design (extended abstract). In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing (STOC)*, 1999, pp. 129–140.

28.9.2 Coordination Mechanisms, Price of Anarchy and Price of Stability

Investigators: Giorgos Christodoulou, Evangelia Pyrga, and Rob van Stee

In [4], we introduce an alternative framework, *Coordination Mechanisms*, in order to improve the performance in systems with independent selfish and non-colluding agents. The quality of a coordination mechanism is measured by its price of anarchy—the worst-case performance of a Nash equilibrium over the (centrally controlled) social optimum. A Nash equilibrium is a situation where no player can change his strategy unilaterally and benefit. Taking a less pessimistic approach, it is also interesting to study the price of stability, which compares the *best* Nash equilibrium to the social optimum. In [3], we study the design of *truthful coordination mechanisms* for task allocation problems. We come up with upper and lower bounds for centrally controlled, as well as for distributed truthful coordination mechanisms.

Another problem we have studied is contention resolution [7]. This arises in various applications of multi-access communications, one example being the celebrated ALOHA protocol [1]. We consider a set of users that all want to transmit one packet via a wireless channel. If more than one users attempt transmission at the same time, a *collision* occurs and they all have to try again later. The question that arises is how to coordinate the users in a decentralized and fair manner, such that they can all succeed within short time. Assuming though that users are selfish players, interested in minimizing their own time of successful transmission, they might end up playing equilibrium strategies that are far from the transmission policies of an efficient protocol. Fiat et al. [10] showed the existence of a *blocking equilibrium*, in which all players transmit constantly, thus blocking the channel indefinitely, and proved that even non-blocking equilibria strategies can be arbitrarily inefficient. Exploiting the blocking equilibrium they gave a protocol (in equilibrium), such that all users succeed (w.h.p.) in linear expected time. However, if we assume that there is an additional cost for every transmission (modelling for instance energy consumption) then the protocol proposed

in [10] no longer forms an equilibrium. We propose two protocols [7] that can work with positive transmission costs, for two different channel feedback structures. Our protocols form $(1 + o(1))$-equilibria, in the sense that if a player that follows the protocol (deviates) has expected cost C (C'), then $C \leq (1 + o(1))C'$.

Scheduling on Related Machines. In scheduling problems, the atomic players are the jobs, and the delay of a job is the completion time of the machine running it, also called the load of this machine. The social goal is to minimize the maximum delay of any job, while the selfish goal of each job is to minimize its own delay, that is, the delay of the machine running it.

We consider scheduling on uniformly related machines [9]. While previous studies either consider identical speed machines or an arbitrary number of speeds, focusing on the number of machines as a parameter, we consider the situation in which the number of different speeds is small. We reveal a linear dependence between the number of speeds and the POA. For a set of machine of at most p speeds, the POA turns out to be exactly $p + 1$. The growth of the POA for large numbers of related machines is therefore a direct result of the large number of potential speeds. We further consider a well known structure of processors, where all machines are of the same speed except for one possibly faster machine. We investigate the POA as a function of both the speed ratio between the fastest machine and the number of slow machines.

28.9.3 Properties of Nash Equilibria

Investigators: Giorgos Christodoulou and Chien-Chung Huang

Bayesian Combinatorial Auctions. In [6] we study the following Bayesian setting: m items are sold to n selfish bidders in m independent second-price auctions. Each bidder has a *private* valuation function that expresses complex preferences over *all* subsets of items. Bidders only have *beliefs* about the valuation functions of the other bidders, in the form of probability distributions. The objective is to allocate the items to the bidders in a way that provides a good approximation to the optimal social welfare value. We show that if bidders have sub-modular valuation functions, then every Bayesian Nash equilibrium of the resulting game provides a 2-approximation to the optimal social welfare. Moreover, we show that in the full-information game a pure Nash always exists and can be found in time that is polynomial in both m and n.

Price of Anarchy and Stability of ϵ-Nash Equilibria. In [5] we study the performance of approximate Nash equilibria for congestion games with polynomial latency functions. We consider how much the price of anarchy worsens and how much the price of stability improves as a function of the approximation factor ϵ. We give almost tight upper and lower bounds for both the price of anarchy and the price of stability for atomic and non-atomic congestion games. Our results not only encompass and generalize the existing results of exact equilibria to ϵ-Nash equilibria, but they also provide a unified approach which reveals the common threads of the atomic and non-atomic price of anarchy results. By expanding the spectrum, we also cast the existing results in a new light. For example, the Pigou network of two parallel

links, which gives tight results for exact Nash equilibria of selfish routing, remains tight for the price of stability of ϵ-Nash equilibria but not for the price of anarchy.

Nash Equilibria in Flow Games. In an atomic splittable flow game a set of players tries to route their flows from a (set of) source(s) to a (set of) destination(s). The flow each of them controls is non-negligible. They can split their flows and choose their routing strategies freely. Each edge in the given network has a cost function of flow volume. A player incurs a cost on an edge, which is a product of his own flow and the delay determined by the total flow of all players. Players, naturally selfish, want to minimize their own costs and disregard others' welfare.

The atomic splittable flow games capture several real world situations. For instance, in transportation, a player can be regarded as a shipping company that aims to minimize the average delay time of its cargoes. We can also represent the Internet as the network, where each player is a manager of an overlay network. He seeks to route the traffic he controls in the most efficient way.

It is natural to inquire on the structures of *Nash equilibrium* in these games. Cominetti, Correa, and Stier-Moses [8] asked whether Nash equilibria are unique in atomic splittable flow games. In [2], we show that there exist examples in which multiple equilibria exist. More importantly, we show that these examples are topologically minimal by giving a complete characterization of the class of network topologies for which unique equilibria exist. Both of our uniqueness and multiplicity results are based on a new characterization of two classes of graphs in terms of sets of circulations.

References

[1] N. Abramson. The aloha system—another alternative for computer communications. In *Proceedings of Fall Joint Computer Conference, AFIPS Conference*, 1970, pp. 295–298.

- [2] U. Bhaskar, L. Fleischer, D. Hoy, and C.-C. Huang. Equilibria of atomic flow games are not unique. In C. Mathieu, ed., *20th Annual ACM-SIAM Symposium on Discrete Algorithms*, New York, U.S.A., 2009, pp. 748–757. Society for Industrial and Applied Mathematics.
- [3] G. Christodoulou, L. Gourves, and F. Pascual. Scheduling selfish tasks: About the performance of truthful algorithms. In G. Lin, ed., *Computing and Combinatorics, 13th Annual International Conference, COCOON 2007*, Banff, Canada, 2007, *LNCS 4598*, pp. 187–197. Springer.
- [4] G. Christodoulou, E. Koutsoupias, and A. Nanavati. Coordination mechanisms. *Theoretical Computer Science*, 2009.

[5] G. Christodoulou, E. Koutsoupias, and P. G. Spirakis. On the performance of approximate equilibria in congestion games. Submitted.

- [6] G. Christodoulou, A. Kovács, and M. Schapira. Bayesian combinatorial auctions. In L. Aceto, I. Damgård, L. A. Goldberg, M. M. Halldórsson, A. Ingólfsdóttir, and I. Walukiewicz, eds., *Automata, Languages and Programming, 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7-11, 2008, Proceedings, Part I*, Reykjavik, Iceland, 2008, pp. 820–832. Springer.

[7] G. Christodoulou and E. Pyrga. Contention resolution under selfishness. Manuscript, 2009.

[8] R. Cominetti, J. R. Correa, and N. E. Stier-Moses. The impact of oligopolistic competition in networks. *Operation Research*, 2008. to appear, extended abstract entitled "Network games with atomic players" appeared in ICALP 06.

- [9] L. Epstein and R. van Stee. The price of anarchy on uniformly related machines revisited. In B. Monien and U.-P. Schroeder, eds., *Algorithmic Game Theory, First International Symposium, SAGT 2008*, Paderborn, Germany, 2008, *LNCS 4997*, pp. 46–57. Springer.

[10] A. Fiat, Y. Mansour, and U. Nadav. Efficient contention resolution protocols for selfish agents. In *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, 2007, pp. 179–188. Society for Industrial and Applied Mathematics.

28.9.4 Online Algorithms

Investigators: Spyros Angelopoulos, Ho-Leung Chan, Nicole Megow, Pascal Schweitzer, and Rob van Stee

Stochastic Online Scheduling with Precedence Constraints. We consider the preemptive and non-preemptive problems of scheduling jobs with precedence constraints on parallel machines with the objective to minimize the sum of (weighted) completion times. We investigate an online model in which the scheduler learns about a job when all its predecessors have completed. For scheduling on a single machine, we present in [16] online algorithms with matching lower and upper bounds of $\Theta(n)$ and $\Theta(\sqrt{n})$ for jobs with general and equal weights, respectively. We also derive corresponding results on parallel identical machines.

Our result for arbitrary job weights holds even in the more general stochastic online scheduling model where, in addition to the limited information about the job set, processing times are uncertain. For a large class of processing time distributions, we derive also an improved performance guarantee if weights are equal.

Online Unit Clustering. We continue the study of the online unit clustering problem [13], introduced by Chan and Zarrabi-Zadeh [10]. We design a deterministic algorithm with a competitive ratio of 7/4 for the one-dimensional case. This is the first deterministic algorithm that beats the bound of 2. It also has a better competitive ratio than the previous randomized algorithms. Moreover, we provide the first non-trivial deterministic lower bound (1.6), improve the randomized lower bound to 1.5, and prove the first lower bounds for higher dimensions.

We also study several variants and generalizations of the online unit clustering problem, which are inspired by variants of packing and scheduling problems in the literature [12]. In particular, we present optimal online algorithms for clustering with rejection, weighted clustering, and clustering with temporary request points. In addition, we present algorithms and lower bounds for clustering with cardinality constraints and clustering with resource augmentation.

The Online Steiner Tree Problem in Directed Graphs. Steiner tree problems occupy a central place in the area of approximation and online algorithms. Apart from their theoretical significance, Steiner tree formulations are also useful in modeling multicast communication in networks. Most of the existing research on Steiner trees assumes undirected graphs. In contrast, however, typical communication networks consist of links asymmetric in the characteristics of their antiparallel links (such as bandwidth and latency), and thus are better modeled by directed graphs.

The above considerations motivated the definition of *edge asymmetry* α, originally due to Ramanathan [17], as the maximum ratio of the weight of antiparallel links in the graph. Ramanathan's work is focused on approximation algorithms (i.e., off-line algorithms). Faloutsos *et gal.* [14], followed by Angelopoulos [1] considered the online variant of the Steiner tree problem in graphs of bounded asymmetry.

In recent work [2], we improved the upper bound on the competitive ratio of a simple greedy algorithm to $O(\min\left\{\max\left\{\alpha\frac{\log k}{\log\alpha}, \alpha\frac{\log k}{\log\log k}\right\}, k\right\})$, where k is the number of terminals. Since [14] and [1] imply a lower bound of $\Omega\left(\min\left\{\max\left\{\alpha\frac{\log k}{\log\alpha}, \alpha\frac{\log k}{\log\log k}\right\}, k^{1-\epsilon}\right\}\right)$ for all algorithms, this bound is near-tight. The analysis is based on identifying "hard" input graphs (i.e., graphs on which the algorithm does not perform well), and which, surprisingly, have a relatively simple structure.

Bijective and Average Analysis of Online Algorithms. It has long been known that for some fundamental online problems, competitive analysis is not consistent with empirical evaluation. The most notable example is the paging problem: there exists a vast class of algorithms, ranging from extremely naive and inefficient strategies (such as Flush-When-Full) to strategies of excellent performance in practice (such as Least-Recently-Used (LRU)) all of which attain the same competitive ratio. A similar situation arises in the list update problem: in particular, under a natural cost formulation, all algorithms have the same asymptotic, non-constant competitive ratio.

In order to bridge this gap between theoretical analysis and empirical performance, Angelopoulos, Dorrigiv and López-Ortiz [3] introduced *bijective analysis* as a natural and intuitive framework for comparing the performance of paging strategies. Given two algorithms A and B, denote by $A(\sigma)$ and $B(\sigma)$ the cost incurred by the algorithms on a request sequence σ. Let also $\mathcal{I}_n$ denote the set of all request sequences of size n. We say that A is *no worse* than B according to bijective analysis, if for all $n \geq n_0$ (for some constant n_0), there exists a permutation $\pi : \mathcal{I}_n \to \mathcal{I}_n$ such that $A(\sigma) \leq B(\pi(\sigma))$. A relaxed version of bijective analysis is *average analysis* in which one compares the average cost incurred by two algorithms over all request sequences of the same size.

In [4] the same authors extended the results of [3] to the list update problem. More specifically [4] shows that any two (natural) algorithms for list update are equivalent according to bijective analysis. The central result in this work shows that the Move-to-Front algorithm (MTF) is superior to all other algorithms according to average analysis, once locality of reference is considered. To our knowledge, this was the first study of the effect of locality of reference in list update.

In [5], Angelopoulos and Schweitzer extended the results of [3] and [4] using the more powerful technique of bijective analysis. More precisely, we showed that LRU and MTF are the unique optimal online algorithms at the presence of locality of reference. This establishes a theoretical justification of the superiority of these two algorithms, in a strong sense.

Weighted Flow Time. We considered the classic online scheduling problem of minimizing the total weighted flow time on a single machine with preemptions. Here, each job j has an arbitrary arrival time r_j, weight w_j and size p_j, and given a schedule its flow time is defined

as the duration of time since its arrival until it completes its service requirement. The first non-trivial algorithms with poly-logarithmic competitive ratio for this problem were obtained relatively recently [11, 8], and it was widely believed that the problem admits a constant factor competitive algorithm. In [6], we showed an $\omega(1)$ lower bound on the competitive ratio of any deterministic online algorithm. Our result was based on a gap amplification technique for online algorithms. Starting with a trivial lower bound of 1, we gave a procedure to improve the lower bound sequentially, while ensuring at each step that the size of the instance increases relatively modestly. Future work includes closing the gap between the upper and lower bounds and investigating the possibility of O(1)-competitive randomized algorithms.

Energy Efficient Scheduling. Energy consumption has become a key issue in the design of microprocessors. Major chip manufacturers, such as Intel, AMD and IBM, now produce chips with dynamically scalable speeds, and produce associated software that enables an operating system to manage power by scaling processor speed. Within the last few years there has been a significant amount of research on the scheduling problems that arise in this setting (see, e.g., [15] for a survey).

All of the theoretical speed scaling research to date has assumed that the power function, which expresses the power consumption P as a function of the processor speed s, is of the form $P = s^\alpha$, where $\alpha > 1$ is some constant. Motivated in part by technological advances, in [7], we initiated the study of speed scaling with arbitrary power functions. We considered the problem of minimizing the total flow time plus energy. Our main result was a $(3 + \epsilon)$-competitive algorithm for this problem, that holds for essentially any power function. We also gave a $(2 + \epsilon)$-competitive algorithm for the objective of fractional weighted flow plus energy. Even for power functions of the form s^α, it was not previously known how to obtain competitiveness independent of α for these problems. We also introduced a model of allowable speeds that generalizes all known models in the literature.

We then studied online non-clairvoyant speed scaling to minimize total flow time plus energy. By non-clairvoyant, it means the size of a job is not known until it is completed. In [9], we first considered the traditional model where the power function is $P(s) = s^\alpha$. We gave a non-clairvoyant algorithm that was shown to be $O(\alpha^3)$-competitive. We then showed an $\Omega(\alpha^{1/3-\epsilon})$ lower bound on the competitive ratio of any non-clairvoyant algorithm. We also showed that there are power functions for which no non-clairvoyant algorithm can be $O(1)$-competitive. Note that modern processors usually have multicores and large scale server farms can have millions of processors. The most interesting future work is to extend the study of energy efficient scheduling to the multiprocessor setting.

References

- [1] S. Angelopoulos. Improved bounds for the online steiner tree problem in graphs of bounded edge-asymmetry. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, New Orleans, Louisiana, USA, 2007, pp. 248–257. SIAM.
- [2] S. Angelopoulos. A near-tight bound for the online steiner tree problem in graphs of bounded asymmetry. In *16th Annual European Symposium on Algorithms (ESA 2008)*, Karlsruhe, 2008, *LNCS 5193*, pp. 76–87. Springer.

- [3] S. Angelopoulos, R. Dorrigiv, and A. López-Ortiz. On the separation and equivalence of paging strategies. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, New Orleans, Louisiana, USA, 2007, pp. 229–237. SIAM.

- [4] S. Angelopoulos, R. Dorrigiv, and A. Lopez-Ortiz. List update with locality of reference. In *8th Latin American Symposium on Theoretical Informatics (LATIN 2008)*, Buzios, Brasil, 2008, *LNCS 4957*, pp. 399–410. Springer.

- [5] S. Angelopoulos and P. Schweitzer. Paging and list update under bijective analysis. In C. Mathieu, ed., *20th ACM-SIAM Symposium on Discrete Algorithms (SODA 2009)*, New York, 2009, pp. 1136–1145. ACM press.

- [6] N. Bansal and H.-L. Chan. Weighted flow time does not admit o(1)-competitive algorithms. In C. Mathieu, ed., *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New York, USA, 2009, pp. 1238–1244. ACM Press.

- [7] N. Bansal, H.-L. Chan, and K. Pruhs. Speed scaling with an arbitrary power function. In C. Mathieu, ed., *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New York, USA, 2009, pp. 693–701. ACM Press.

[8] N. Bansal and K. Dhamdhere. Minimizing weighted flow time. *ACM Transactions on Algorithms, (*SODA 2002 and 2003 special issue*)*, 3(4), 2007.

- [9] H.-L. Chan, J. Edmonds, T.-W. Lam, L.-K. Lee, A. Marchetti-Spaccamela, and K. Pruhs. Nonclairvoyant speed scaling for flow and energy. In S. Albers and J.-Y. Marion, eds., *Proceedings of the 26th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Freiburg, Germany, 2009, pp. 255–264. IBFI.

[10] T. M. Chan and H. Zarrabi-Zadeh. A randomized algorithm for onine unit clustering. In *Proc. 4th Workshop on Approximation and Online Algorithms (WAOA 2006)*, 2007, *LNCS 4368*, pp. 121–131. To appear in *Theory of Computing Systems* (doi:10.1007/s00224-007-9085-7).

[11] C. Chekuri, S. Khanna, and A. Zhu. Algorithms for minimizing weighted flow time. In *Proceedings on 33rd Annual ACM Symposium on Theory of Computing (STOC)*, 2001, pp. 84–93.

- [12] L. Epstein, A. Levin, and R. van Stee. Online unit clustering: Variations on a theme. *Theoretical Computer Science*, 407(1-3):85–96, 2008.

- [13] L. Epstein and R. van Stee. On the online unit clustering problem. *ACM Transactions on Algorithms*, 2009. To appear.

[14] M. Faloutsos, R.Pankaj, and K. C. Sevcik. The effect of asymmetry on the on-line multicast routing problem. *Int. J. Foundations of Computer Science*, 13(6):889–910, 2002.

[15] S. Irani and K. R. Pruhs. Algorithmic problems in power management. *SIGACT News*, 36(2):63–76, 2005.

[16] N. Megow and T. Vredeveld. Stochastic online scheduling with precedence constraints. Submitted, 2008.

[17] S. Ramanathan. Multicast tree generation in networks with asymmetric links. *IEEE/ACM Transactions on Networking*, 4(4):558–568, 1996.

28.10 Information Retrieval

Coordinator: Hannah Bast

Making large amounts of information accessible in a way such that sought-after items can be located quickly and intuitively, is still a problem far from being solved. We do research on all facets of information retrieval, with an emphasis on reconciling complex functionality with high scalability. Or, in simpler words: *we want to make intelligent search fast.*

Our work encompasses the whole chain: from the identification, formalization and theoretical analysis of core problems, over carefully engineered implementation and extensive experimentation, to the development of full-fledged usable systems.

The period covered by the biennial report has been special in several respects:

(1) I am on leave from the Max Planck Institute for Informatics since April 2008, when I started an appointment as a visiting scientist at Google Zurich. This appointment was originally planned for half a year, but my project there went so well (and needed me) that I prolonged it to a full year.

(2) During my stay at Google I received three offers: for a full professorship (W3) at the University of Mainz, for a full professorship at the University of Freiburg (W3), and for a full-time research position at Google. I will accept the offer from Freiburg.

(3) Due to the combination of being away from the institute for an extended period of time and then knowing that I would leave Max Planck Institute for Informatics in 2009 for one position or the other, I stopped hiring people in my group and since 2008 was left with only a single PhD student: Marjan Celikik. (Debapriyo Majumdar and Ingmar Weber successfully defended their thesis and left the institute at the end of 2007.)

(4) The work of our group received a number of prestigious awards in the period covered by this biennial report: the Meyer-Struckmann-Science Award and the Alcatel-Lucent Award for Technical Communication. I also received two awards for my work on route planning (together with Stefan Funke).

So, all in all, it has been a fairly successful period, especially for myself :-)

28.10.1 Semantic Search

Investigators: Hannah Bast, Alexandru Chitea, Fabian Suchanek, and Ingmar Weber

We developed ESTER, a modular and highly efficient system for combined full-text and ontology search. ESTER builds on a query engine that supports two basic operations: prefix search and join. Both of these can be implemented very efficiently with a compact index, yet in combination provide powerful querying capabilities. The implementation is an extension of our CompleteSearch technology, described at length in the last biennial report.

ESTER can answer basic SPARQL graph-pattern queries on the ontology by reducing them to a small number of these two basic operations. ESTER further supports a natural blend of such semantic queries with ordinary full-text queries. Moreover, the prefix search operation allows for a fully interactive and proactive user interface, which after every keystroke suggests to the user possible semantic interpretations of his or her query, and speculatively executes the most likely of these interpretations.

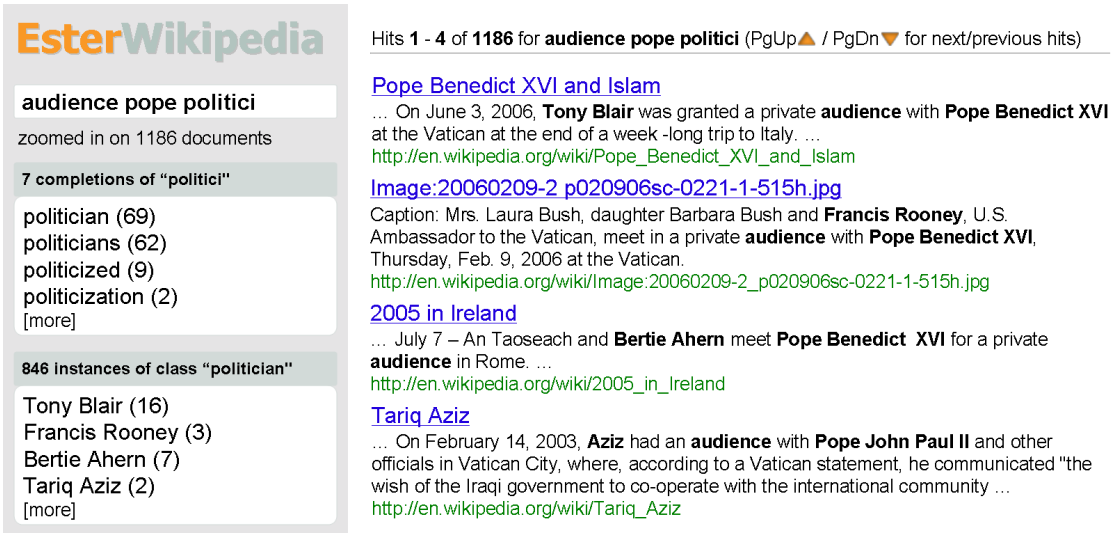


Figure 28.15: A screenshot of our semantic search engine ESTER. You can try it out live under `http://search.mpi-inf.mpg.de/ester`.

As a proof of concept, we applied ESTER to the English Wikipedia, which contains about 3 million documents, combined with the recent YAGO ontology, which contains about 2.5 million facts. For a variety of complex queries, ESTER achieves worst-case query processing times of a fraction of a second, on a single machine, with an index size of about 4 GB.

References

- [1] H. Bast, A. Chitea, F. Suchanek, and I. Weber. Ester: Efficient search in text, entities, and relations. In C. Clarke, N. Fuhr, and N. Kando, eds., *30th International Conference on Research and Development in Information Retrieval (SIGIR'07)*, Amsterdam, Netherlands, 2007, pp. 671–678. ACM.
- [2] H. Bast, F. M. Suchanek, and I. Weber. Semantic full-text search with ester: Scalable, easy, fast. In *8th International Conference on Data Mining (ICDM'08)*, 2008, pp. 959–962.

28.10.2 Error-tolerant Search

Investigators: Hannah Bast and Marjan Celikik

We introduced the following spelling variants clustering problem: Given a list of distinct words, called lexicon, compute (possibly overlapping) clusters of words which are spelling variants of each other. This problem naturally arises in the context of error-tolerant full-text search of the following kind: For a given query, return not only documents matching the query words exactly but also those matching their spelling variants. This is the inverse of the well-known "Did you mean: ... ?" web search engine feature, where the error tolerance is on the side of the query, and not on the side of the documents.

We combined various ideas from the the large body of literature on approximate string searching and spelling correction problems to a new algorithm for the spelling variants

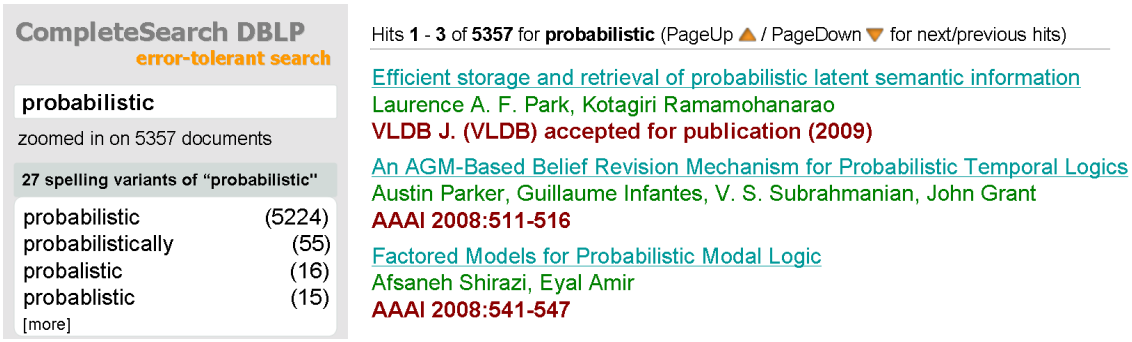


Figure 28.16: Our error-tolerant search in action. You can try it out live on `http://search.mpi-inf.mpg.de/dblp`

clustering problem that is both accurate and very efficient in time and space. Our largest lexicon, containing roughly 25 million words, can be processed in about 18 minutes on a standard PC using 10 MB of additional space. This beats the previously best scheme by a factor of two in running time and by a factor of more than ten in space usage. We have integrated our algorithms into our CompleteSearch engine in a way that achieves error-tolerant search without significant blowup in neither index size nor query processing time.

Ours is the first work on efficiently achieving error-tolerance on the document side. In particular, we compare ourselves to the popular Lucene search engine, which implements this feature via straightforward query expansion. Our solution is faster by more than an order of magnitude, at the price of only a small increase in index size.

References

- [1] M. Celikik and H. Bast. Fast error-tolerant search on very large texts. In D. Shin, ed., *The 24th Annual ACM Symposium on Applied Computing*, Honolulu, Hawaii, USA, 2009, *PROCEEDINGS OF THE 2009 ACM SYMPOSIUM ON APPLIED COMPUTING*, vol. 104092, pp. 1724–1731. ACM.

28.10.3 Snippet Generation

Investigators: Hannah Bast, Marjan Celikik, and Gabriel Manolache

Ranked result lists with query-dependent snippets have become state of the art in text search. They are typically implemented by searching, at query time, for occurrences of the query words in the top-ranked documents. This *document-based* approach has three inherent problems: (i) when a document is indexed by terms which it does not contain literally (e.g., related words or spelling variants), localization of the corresponding snippets becomes problematic; (ii) each query operator (e.g., phrase or proximity search) has to be implemented twice, on the index side in order to compute the correct result set, and on the snippet generation side to generate the appropriate snippets; and (iii) in a worst case, the whole document needs to be scanned for occurrences of the query words, which is problematic for very long documents.

Figure 28.17: Some variants of snippet generation which our new algorithm can deal with efficiently.

We have introduced a new *index-based* method that localizes snippets by information solely computed from the index, and that overcomes all three problems. Unlike previous index-based methods, we show how to achieve this at essentially no extra cost in query processing time, by a technique we call *query rotation*. We also show how our index-based method allows the caching of individual segments instead of complete documents, which enables a significantly larger cache hit ratio as compared to the document-based approach. We have fully integrated our implementation with the CompleteSearch engine.

References

- [1] M. Celikik, H. Bast, and G. Manolache. Efficient index-based snippet generation. Rejected paper. Under revision., 2009.

28.10.4 Index Construction

Investigators: Hannah Bast and Marjan Celikik

The core of our CompleteSearch technology is the so-called hybrid or HYB index. The HYB index allows fast context-sensitive prefix search, which in turn is the basis for the efficient support of many other advanced query types (for example, the error-tolerant from Section 28.10.7).

We have now shown that the HYB index is not only more powerful than an inverted index, but can also be constructed in only half the time. Our construction algorithm is truly single-pass in that every posting (word-in-document occurrence) is touched only once in the whole construction. A critical of the construction of a HYB index is the partitioning of the vocabulary into ranges: each range corresponds to a block of postings (all postings with a word in that range), and these blocks should be of similar size. We show how to find a good block partitioning by appropriate random sampling from the documents.

Our algorithm has been carefully engineered and implemented, with special attention paid to IO- and cache-efficiency. Our factor-2 improvement is achieved in comparison to the state-of-the-art (and highly optimized) index construction from Zettair [2].

References

- [1] M. Celikik and H. Bast. Fast construction of a halfinverted index. 2009.

[2] S. Heinz and J. Zobel. Efficient single-pass index construction for text databases. *JASIST*, 54(8):713–729, 2003.

28.10.5 IO-Efficient Faceted Search

Investigators: Omid Amini, Hannah Bast, Hubert Chan, and Andreas Karrenbauer

In faceted search, we are given a collection of text documents, where additionally each document is annotated with a number of labels which are organized into categories or *facets*. For example, consider a collection of computer science articles, with each document annotated by its authors, the name of the conference where it was published and the year of publication. A faceted search interface allows the user to alternatively search by keyword or browse by these categories; for an example see `http://dblp.mpi-inf.mpg.de/dblp`.

In the previous report we have described an implementation of faceted search based on our CompleteSearch technology. This works like a charm for most collections but hits some performance limits when the collection becomes very large (tens of millions of documents).

In this more theoretical work we modelled the problem as follows: Preprocess a given set of n labeled objects such that for a given subset of objects, called a query, we can quickly compute the multiset of labels of the queried objects. Access to labels is in blocks of size B and labels may be stored redundantly in multiple locations. We study the trade-off between the amount of redundancy when storing the labels and the number of block accesses required to answer a query.

We proved various upper and lower bounds on this trade-off, under various assumptions on the query distribution. Our main result assumes that queries come from a fixed set and that each query is a random subset; both of these are realistic in typical applications of faceted search. Under these assumptions, we show that an optimal number of block accesses can be achieved with space consumption for the labels being an arbitrarily small fraction ε of the amount of space needed to store the queries, provided that n is sufficiently large. In our experiments, we provide concrete numbers, for example, we show that $\varepsilon = 1/4$ can be achieved for $n \geq 4.5 \cdot 10^6 \cdot B$.

References

- [1] O. Amini, H. Bast, H. Chan, and A. Karrenbauer. Space-time trade-offs for large-scale faceted search. 2009.

28.10.6 Efficient Top-k Retrieval

Investigators: Hannah Bast, Debapriyo Majumdar, Ralf Schenkel, and Martin Theobald

This work was done in cooperation with the Department for Databases and Information Systems (D5), and is described in more detail in Section 31.

28.10.7 Efficient Large-scale 3D-Shape Retrieval

Investigators: Hannah Bast, Joachim Giesen, and Waqar Saleem

This work was done in cooperation with the Department for Computer Graphics (D4), and is described in more detail in Section 30.

28.11 Partner Group on Approximation Algorithms

Coordinator: Naveen Garg

The Algorithms Group at IIT Delhi, India was designated a Partner-Group of Max Planck Institute for Informatics for research in the area of "Approximation Algorithms". The group started its operations in July 2005.

28.11.1 Scheduling to Minimize Flow Time

Investigators: Naveen Garg, Amit Kumar, V. N. Muralidhara, Jivitej S. Chadha, and Vinayaka Pandit

We consider the problem of scheduling jobs on multiple machines so as to minimize the total flow time. The flow time of a job is the total time it spends in the system and equals the difference between its completion time and release time.

Unrelated Machines

A job j has a processing time p_{ij} on machine i; this is the most general model of processing times and is known as the *unrelated machines model*.

In [3] we consider a special case of scheduling on unrelated machines. The machines are identical except that each job can be assigned only to a specified subset of the machines. We show that no online algorithm can have a bounded competitive ratio. We provide an $O(\log P)$-approximation algorithm by modifying the single-source unsplittable flow algorithm of Dinitz et.al.[2]. Here P is the ratio of the maximum to the minimum processing times. We also establish an $\Omega(\log P)$-integrality gap for our LP-relaxation and use this to show an $\Omega(\log P/\log\log P)$ lower bound on the approximability of the problem.

For the unrelated machines setting, we introduce a notion of (α, β) variability to capture settings where processing times of jobs on machines are not completely arbitrary. We say that processing times have an (α, β)-*variability* if the processing time of job j on machine i can be expressed as $p_{ij} = a_j \cdot b_{ij} \cdot s_i$ where $b_{ij} \in B, |B| = \beta$ and $1 \leq a_j \leq \alpha$. Our main result in [4] is a simple $O(\beta \log \alpha)$ approximation for this setting. As special cases, we get

(a) an $O(k)$ approximation when there are only k different processing times.

(b) an $O(\log P)$-approximation if each job can only go on a specified subset of machines, but has the same processing requirement on each such machine. Further, the machines can have different speeds. Here P is the ratio of the largest to the smallest processing requirement.

(c) an $O(\epsilon^{-1} \log \epsilon^{-1})$- approximation algorithm for unrelated machines if we assume that our algorithm has machines which are an $(1 + \epsilon)$-factor faster than the optimum algorithm's machines.

We also extend the hardness results to the problem of minimizing flow time on parallel machines. We show that the problem cannot be approximated to within $\Omega(\log^{1-\epsilon} P)$ for any $\epsilon > 0$.

Unrelated Machines with Speed Augmentation

We consider the online problem of scheduling jobs on unrelated machines so as to minimize the total weighted flow time. This problem has an unbounded competitive ratio even for very restricted settings. In [1] we show that if we allow the machines of the online algorithm to have ϵ more speed than those of the offline algorithm then we can get an $O((1+\epsilon^{-1})^2)$-competitive algorithm.

Our algorithm schedules jobs preemptively but without migration. However, we compare our solution to an offline algorithm which allows migration. Our analysis uses a potential function argument which can also be extended to give a simpler and better proof of the randomized immediate dispatch algorithm of Chekuri-Goel-Khanna-Kumar for minimizing average flow time on parallel machines.

Order Scheduling

In [5] we consider scheduling problems in which a job consists of components of different types to be processed on m machines. Each machine is capable of processing components of a single type. Different components of a job are independent and can be processed in parallel on different machines. A job is considered as completed only when all its components have been completed. We study both completion and flow time aspects of such problems.

We show both lower bounds and upper bounds for the completion time problem. We first show that even the unweighted completion time with single release date is MAX-SNP hard. We give an approximation algorithm based on linear programming which has an approximation ratio of 3 for weighted completion time with multiple release dates. We give online algorithms for the weighted completion time which are constant factor competitive.

For the flow time, we give only lower bounds in both the offline and online settings. We show that it is NP-hard to approximate flow time within $\Omega(\log m)$ in the offline setting. We show that no online algorithm for the flow time can have a competitive ratio better than $\Omega(\sqrt{m})$.

References

[1] J. S. Chadha, N. Garg, A. Kumar, and V. N. Muralidhara. A competitive algorithm for minimizing weighted flow time on unrelated machines with speed augmentation. In *STOC*, 2009.

[2] Y. Dinitz, N. Garg, and M. X. Goemans. On the single-source unsplittable flow problem. *Combinatorica*, 19(1):17–41, 1999.

[3] N. Garg and A. Kumar. Minimizing average flow-time : Upper and lower bounds. In *FOCS*, 2007, pp. 603–613.

[4] N. Garg, A. Kumar, and V. N. Muralidhara. Minimizing total flow-time: The unrelated case. In *ISAAC*, 2008, pp. 424–435.

[5] N. Garg, A. Kumar, and V. Pandit. Order scheduling models: Hardness and algorithms. In *FSTTCS*, 2007, pp. 96–107.

28.11.2 Stochastic Analysis of Online Algorithms

Investigators: Naveen Garg, Anupam Gupta, Stefano Leonardi, and Piotr Sankowski

The study of online algorithms has been an extremely popular and successful program in the area of algorithms. This has mainly focused on *competitive analysis*, where the performance of an online algorithm (that knows nothing about the future) is compared to the optimal solution built with hindsight. This model has led to cleanly defined problems, and strong upper and lower bounds on the competitive ratio are known for most problems of interest. There are, however, shortcomings to using competitive analysis: the biggest objection being that the strict definition of competitive ratio does not allow us to make fine-grained distinctions between algorithms.

Over the years, these drawbacks to the competitive analysis framework have caused researchers to try and weaken the rigid competitive analysis framework, and return to variants of the fundamental question: *Can we do better if we are given access to the input distribution?* While the situation is fairly-well understood for classical online problems like paging and k-server, we still know almost nothing for, say, online Steiner tree in this model. In fact, the starting point of our investigation is the following basic question:

> *Can we beat the* $\Theta(\log n)$ *bound known for online Steiner tree if at each time instant, the demand vertex is a uniformly random vertex from the graph?*

In [1], we answer this question in the affirmative: we show that if each vertex is an independent draw from some probability distribution $\pi : V \to [0, 1]$, a slight variant of the natural greedy algorithm achieves either an $O(1)$, or an $O(\log \log n)$ performance guarantee, depending on the precise nature of the guarantees desired. Some of these results can be extended to other subadditive problems as well. Furthermore, we show that both assumptions that the input sequence consists of *independent* draws from π, and that π is known to the algorithm are both essential; we show logarithmic lower bounds if either assumption is violated.

References

[1] N. Garg, A. Gupta, S. Leonardi, and P. Sankowski. Stochastic analyses for online combinatorial optimization problems. In *SODA*, 2008, pp. 942–951.

28.11.3 Stochastic Network Design

Investigators: Anupam Gupta, Amit Kumar and MohammadTaghi Hajiaghayi

While the Steiner tree problem has been well studied in the model of two stage stochastic optimization with recourse, with several different solutions and extensions to the multistage case, all these algorithms work only in the case when all the edge-costs increase by a uniform factor. When each edge-cost is allowed to increase by a different factor, nothing was previously known.

In [1] we show that this problem admits a poly-logarithmic approximation guarantee; moreover, it is as hard as the cost-distance problem, for which we have a $\Omega(\log \log n)$ hardness. Also, we show that if the inflation is allowed to vary over scenarios, the problem becomes as hard as Label-cover. Finally, we give a new linear-programming relaxation of the multi-commodity cost-distance problem.

In [2] we consider the stochastic Steiner forest problem: suppose we were given a collection of Steiner forest instances, and were guaranteed that a random one of these instances would appear tomorrow; moreover, the cost of edges tomorrow will be λ times the cost of edges today. Which edges should we buy today so that we can extend it to a solution for the instance arriving tomorrow, to minimize the expected total cost? While very general results have been developed for many problems in stochastic discrete optimization over the past years, the approximation status of the stochastic Steiner Forest problem has remained open, with previous works yielding constant-factor approximations only for special cases. We resolve the status of this problem by giving a constant-factor primal-dual based approximation algorithm.

References

[1] A. Gupta, M. Hajiaghayi, and A. Kumar. Stochastic steiner tree with non-uniform inflation. In *APPROX-RANDOM*, 2007, pp. 134–148.

[2] A. Gupta and A. Kumar. A constant factor approximation for stochastic steiner forest. In *Proceedings on 41st Annual ACM Symposium on Theory of Computing (STOC)*, 2009.

28.12 Partner Group on Efficient Graph Algorithms

Coordinator: Telikepalli Kavitha

The Efficient Graph Algorithms Group at Indian Institute of Science, Bangalore was designated as a Partner Group of the Max Planck Institute for Informatics for research in the area of efficient graph algorithms. The group started its operations in March 2008.

28.12.1 Matching Problems in Bipartite Graphs

Investigators: Telikepalli Kavitha and Meghana Nasre

These problems deal with matching a set of *applicants* to a set of *posts* (the two sides of a bipartite graph G) with one-sided preference lists: that is, each applicant ranks a non-empty subset of posts in an order of preference, possibly involving ties and we seek

to match applicants to posts keeping in mind the preference lists of the applicants. This model captures some important real-world problems, including the allocation of graduates to training positions and families to government-owned housing.

There are various criteria for measuring how good a matching is. A natural optimality criterion is *popularity*. A matching M is more popular than matching M' if the number of applicants who prefer M to M' outnumber the number of applicants who prefer M' to M. A matching M^* is popular if there is no matching that is more popular than M^*.

An attractive feature about popularity is that it does not use numerical ranks. However, not every graph admits a popular matching. Abraham et al. [1] studied the popular matching problem and gave efficient algorithms for determining if G admits a popular matching and if so computing one.

We considered an extension of this problem: given that G admits a popular matching, compute an *optimal* popular matching, where the input contains a succinct definition of optimality. For example, rank-maximality, fairness, or min-cost of matched edges can be considered as optimality. We showed an $O(n^2 + m)$ algorithm for computing an optimal popular matching in G (where n is the number of vertices and m is the number of edges) assuming that the preference lists are *strict*. A preliminary version of this work appeared in [3].

Bounded Unpopular Matchings

If the input graph G does not admit a popular matching, the next best alternative would be to ask for a *least unpopular* matching. In [4] McCutchen used two criteria (i) least unpopularity margin and (ii) least unpopularity factor to measure the unpopularity of a matching.

Given any two matchings M and M' in G, let $\phi(M, M')$ be the number of applicants that prefer M to M'. The unpopularity factor of M with respect to M', $\Delta(M, M')$, is defined as $\phi(M', M)/\phi(M, M')$ if $\phi(M, M') > 0$, it is 1 if both $\phi(M, M')$ and $\phi(M', M)$ are 0, otherwise it is ∞. The unpopularity factor of a matching M is defined as $\max_{M'} \Delta(M, M')$.

The unpopularity margin of M with respect to M', $\delta(M, M')$, is defined as $\phi(M', M) - \phi(M, M')$. The unpopularity margin of a matching M is defined as $\max_{M'} \delta(M, M')$. McCutchen showed that computing a matching that minimizes unpopularity factor or unpopularity margin is NP-hard.

We considered efficient approximation algorithms for these problems. We showed that a matching M that achieves an unpopularity factor of 2 can be computed in $O(m\sqrt{n})$ time (where m is the number of edges in G and n is the number of vertices) provided a certain graph H admits a matching that matches all applicants. We also show a sequence of graphs: $H = H_2, H_3, \ldots, H_k$ such that if H_k admits a matching that matches all applicants, then we can compute in $O(km\sqrt{n})$ time a matching M whose unpopularity factor is at most k and whose unpopularity margin is at most $n(1 - \frac{2}{k})$.

We ran our algorithm on random graphs and our simulation results suggest that in random graphs our algorithm terminates in a small constant number of rounds. Thus these graphs admit matchings whose unpopularity factor is a small constant and whose unpopularity margin can be bounded away from n. This is joint work with Chien-Chung Huang and Dimitrios Michail; a preliminary version of this work appeared in [2].

Duplicating/Closing Down Posts

Our next attempt to find an answer when G does not admit a popular matching was the following more general question: given $G = (A \cup \{p_1, \ldots, p_k\}, E)$ and a vector $(c_1, \ldots, c_k)$, is there a set of values $x_1, \ldots, x_k$ such that $1 \leq x_i \leq c_i$ for each i and making x_i copies of post p_i would enable the new graph to admit a popular matching. Similarly, we also considered the question of closing down posts. That is, if we are given a subset P of posts, each of which can be independently closed down or kept open, can we determine $S \subseteq P$ such that $(A \cup \{p_1, \ldots, p_k\} \setminus S, E)$ admits a popular matching.

We show that both these problems are NP-hard. We show that given a graph G, the problem of determining if there is a subset of posts which if duplicated, enable the new graph to admit a popular matching, is NP-hard. This problem is a special case of both the problems in the preceding paragraph.

Popular Mixed Matchings

We also studied the problem of computing a probability distribution over matchings, in other words, a *mixed matching*, that is popular. We showed that unlike popular (pure) matchings, popular mixed matchings are always guaranteed to exist; also such a mixed matching can be computed in polynomial time. This is joint work with Julian Mestre.

References

[1] D. Abraham, R. Irving, T. Kavitha, and K. Mehlhorn. Popular matchings. *SIAM Journal on Computing*, 37(4):1030–1045, 2007.

- [2] C.-C. Huang, T. Kavitha, D. Michail, and M. Nasre. Bounded unpopularity matchings. In *11th Scandinavian Workshop on Algorithm Theory (SWAT)*, Gothenburg, Sweden, 2008, pp. 127–137. Springer.

- [3] T. Kavitha and M. Nasre. Optimal popular matchings, 2008.

[4] R. M. McCutchen. The least-unpopularity-factor and least-unpopularity-margin criteria for matching problems with one-sided preferences. In *Proceedings of the 15th Latin American Symposium on Theoretical Informatics*, 2008, pp. 593–604.

28.12.2 Minimum Co-cycle Basis Problems

Investigator: Telikepalli Kavitha

The problem of computing a minimum cycle basis in a graph G (directed or undirected) is well-studied. Liebchen and Rizzi [2] studied various classes of cycle bases for general graphs; this refined classification of cycle bases was of strong relevance for practical applications and they identified several new variants of the minimum cycle basis problem. More precisely, they showed that for general graphs, computing a minimum cycle basis among certain classes of cycle bases is different from computing a minimum cycle basis among *any* of the other classes.

We studied this question for minimum co-cycle basis problems. The co-cycle space of G is the orthogonal complement of its cycle space. While co-cycles in an undirected graph are simply cuts, a co-cycle in a directed graph corresponds to a vertex partition (S, S^c): an

$\{-1, 0, 1\}$ edge incidence vector C describes this vertex partition where $C(e) = 1$ if e is from S to S^c, $C(e) = -1$ if e is from S^c to S, and $C(e) = 0$ otherwise.

We show that there is a special co-cycle basis in any directed graph G such that this set of co-cycles simultaneously answers most of the minimum co-cycle basis problems in G. This co-cycle basis corresponds to the cuts of a Gomory-Hu tree of the underlying undirected graph of G. It is known that this is a minimum undirected co-cycle basis. We show that this is also a minimum directed co-cycle basis and it is weakly fundamental. It also follows from known results that the matrix of this co-cycle basis is totally unimodular.

It thus follows that this set of co-cycles is a minimum directed co-cycle basis, a minimum undirected co-cycle basis, a minimum integral co-cycle basis, a minimum weakly fundamental co-cycle basis, and a minimum totally unimodular co-cycle basis of G. A preliminary version of this work appeared in [1].

References

- [1] T. Kavitha. On a special co-cycle basis of graphs. In *11th Scandinavian Workshop on Algorithm Theory (SWAT)*, Gothenburg, Sweden, 2008, pp. 343–354. Springer.

[2] C. Liebchen and R. Rizzi. Classes of cycle bases. *Discrete Applied Mathematics*, 155(3):337–355, 2007.

28.13 Academic Activities

28.13.1 Journal Positions

Benjamin Doerr is on the editorial board of

- *Information Processing Letters* (since 2008).

Kurt Mehlhorn is on the editorial board of

- *Computational Geometry: Theory and Applications*,electronic submission, Editor in Chief (since 1990).
- *ACM Transactions on Algorithms*, Editor (since 2004).
- *Japan Journal of Industrial and Applied Mathematics (JJIAM)*, Member of Advisory Board (since 2004).
- *Mathematical Programming Computation*, Member of Advisory Board (since 2008).

28.13.2 Conference and Workshop Positions

Membership in Program Committees

Hannah Bast:

- *Workshop on Dynamic Taxonomies and Faceted Search (FIND 2007)*, Regensburg, Germany, September 2007
- *European Conference on Information Retrieval (ECIR 2007)*, Rome, Italy, April 2007.

- *Research and Development in Information Retrieval (SIGIR 2007)*, Amsterdam, Netherlands, July 2007
- *European Conference on Information Retrieval (ECIR 2008)*, Glasgow, Scotland April 2008.
- *Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, July 2008
- *Web Search and Data Mining (WSDM 2008)*, Stanford, USA, February 2008
- *Symposium on Theoretical Aspects of Computer Science (STACS 2009)*, Freiburg, Germany, February 2009
- *Research and Development in Information Retrieval (SIGIR 2009)*, Boston, USA, July 2009
- *European Symposium on Algorithms (ESA 2009)*, Copenhagen, Denmark, September 2009
- *Using Search Engine Technology for Information Management (USETIM 2009)*, Lyon, France, August 2009

Giorgos Christodoulou:

- *35th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2009)*, Hotel Arnika, Spindleruv Mlyn, Czech Republic, January 2009
- *2nd International Symposium on Algorithmic Game Theory*, Cyprus, October 2009

Benjamin Doerr:

- *Genetic and Evolutionary Computation Conference 2007*, London, UK, July 2007 (Co-Chair of the track "Formal Theory")
- *IEEE Congress on Evolutionary Computation 2007*, Singapore, September 2007
- *International Conference on Computational Intelligence and Security*, Harbin, China, December 2007
- *IEEE Congress on Evolutionary Computation 2008*, Hong Kong, China, June 2008
- *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, July 2008 (Co-Chair of the track "Formal Theory")
- *International Conference on Parallel Problem Solving From Nature 2008*, Dortmund, Germany, September 2008
- *International Conference on Computational Intelligence and Security*, Suzhou, China, December 2008
- *IEEE Congress on Evolutionary Computation 2009*, Trondheim, Norway, May 2009
- *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, July 2009 (Co-Chair of the track "Theory")

Tobias Friedrich:

- *IEEE Congress on Evolutionary Computation 2008*, Hong Kong, China, June 2008

- *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, July 2008
- *International Conference on Parallel Problem Solving From Nature 2008*, Dortmund, Germany, September 2008
- *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, July 2009

Michael Kerber:

- *Symposium on Computational Geometry (Video and Multimedia Presentations)*, Aarhus, Denmark, June 2009

Nicole Megow:

- *Models and Algorithms for Planning and Scheduling Problems*, Abbey Rolduc, The Netherlands, June/July 2009

Kurt Mehlhorn:

- *European Symposium on Algorithms (ESA 2008)*, Karlsruhe, Germany, September 2008 (Program Chair)
- *Member of Senate, Max-Planck-Gesellschaft* (Advisory Board, 2002-2014)
- *Board of Governors, Jacobs University Bremen* (Advisory Board, since 2006)
- *Joint Advisory Board, Carnegie Mellon, Qatar* (Advisory Board, since 2009)

Frank Neumann:

- *Genetic and Evolutionary Computation Conference 2007*, London, UK, July 2007 (Co-Chair of the track "Formal Theory")
- *Workshop "From Biology To Concurrency and back"*, Lisbon, Portugal, September 2007
- *IEEE Congress on Evolutionary Computation 2007*, Singapore, September 2007
- *IEEE Congress on Evolutionary Computation 2008*, Hong Kong, China, June 2008
- *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, July 2008 (Chair of the track "Evolutionary Combinatorial Optimization")
- *Workshop and Summer School on Evolutionary Computing 2008*, Derry, Northern Ireland, August 2008
- *International Conference on Intelligent Computing 2008*, Shanghai, China, September 2008
- *International Conference on Ant Colony Optimization and Swarm Intelligence 2008*, Brussels, Belgium, September 2008
- *International Conference on Parallel Problem Solving From Nature 2008*, Dortmund, Germany, September 2008
- *IEEE Symposium of Computational Intelligence in Multi-Criteria Decision Making 2009*, Nashville, USA, March 2008
- *IEEE Congress on Evolutionary Computation 2009*, Trondheim, Norway, May 2009
- *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, July 2009

28.13.3 Invited Talks and Tutorials

Benjamin Doerr:

- *Quasirandom Rumor Spreading*, Colloquium on Combinatorics, Magdeburg, November 2008.

Khaled Elbassioni:

- *The Negative Cycles Polyhedron and Hardness of Testing Polyhedral Properties*, DIMACS Workshop on Algorithmic Challenges in Optimization, Game Theory and Computer Science: in Memory of Leonid Khachiyan, Rutgers University, NJ, USA, March 2009.
- *On the Hardness of Approximating the Vertex-Centroid of a Polyhedron*, Optimization Days, University of Montreal, Montreal, Canada, May 2009.

Tobias Friedrich:

- *Deterministic random walks and their application to rumor spreading*, MIT Probability Seminar, Department of Mathematics, Massachusetts Institute of Technology, Boston, USA, February 2009.

Michael Hemmer:

- *Project Management of CGAL*, AVACS Winter School on Distributed Software Development Processes, November 2008.
- *Generic Programming using C++ templates*, AVACS Winter School on Distributed Software Development Processes, November 2008.

Kurt Mehlhorn:

- *Root Isolation for Real Polynomials*, Symposium on Exact Geometric Methods, Minneapolis, May 2007.
- *Matching Problems*, International Conference on Combinatorial Optimization and Application (COCOA) 2007, June 2007.
- *Cycle Bases in Graphs*, International Symposium on Mathematical Foundations of Computer Science (MFCS) 2007, September 2007.
- *Assigning Papers to Referees*, International Colloquium on Automata, Languages and Programming (ICALP) 2008, Satellite Conference on Matching, July 2008.
- *Cycle Bases in Graphs*, Summer School in Computer Science, Erice, Italy, September 2008.
- *Algorithm Engineering*, Summer School in Computer Science, Lipari, Italy, August 2008.
- *Assigning Papers to Referees*, Foundations of Genetic Algorithms (FOGA) X, Orlando, Florida, January 2009.
- *Geometric Computing, Now and Then*, The 25th Annual ACM Symposium on Computational Geometry (SoCG), June 2009.

– *Assigning Papers to Referees*, International Colloquium on Automata, Languages and Programming (ICALP) 2009, July 2009.

Frank Neumann:

– *Computational Complexity and Evolutionary Computation*, Tutorial (with T. Jansen), Genetic and Evolutionary Computation Conference 2007, July 2007.
– *Computational Complexity and Evolutionary Computation*, Tutorial (with T. Jansen), Genetic and Evolutionary Computation Conference 2008, July 2008.
– *Computational Complexity of Evolutionary Computation in Combinatorial Optimization*, Tutorial (with C. Witt), 10th International Conference on Parallel Problem Solving from Nature, September 2008.

Michael Sagraloff:

– *Exact Geometric-Topological Analysis and Triangulation of Algebraic Surfaces*, IMAGINARY-Mini-Symposium, DMV-Tagung 2008, Erlangen, September 2008
– *Efficient and Exact Algorithms for Topology Computations of Algebraic Curves and Surfaces*, Algebraic Geometry and Geometric Modeling, Lijang, China, July 2009

28.13.4 Other Academic Activities

– Eric Beberich and Michael Hemmer are on the editorial board of *Computational Geometry Algorithms Library (CGAL)* (since 2007).

– ADFOCS 2007 organized by Ernst Althaus, Joachim Giesen, and Evangelia Pyrga, Saarbrücken, September 2007.
Topics: Learning, Prediction and Games, Stability of Clustering, and Topological Data Analysis.
Speakers:

  * Nicoló Cesa-Bianchi, Universitá degli Studi di Milano
  * Gábor Lugosi, Pompeu Fabra University, Barcelona
  * Hans Ulrich Simon, Ruhr-Universität Bochum
  * Afra Zomorodian, Dartmouth College

– ADFOCS 2008 organized by Khaled Elbassioni and Kurt Mehlhorn, Saarbrücken, August 2008.
Topics: Metric Techniques in Approximation Algorithms, Applications of Cost-Sharing Methods to Hard Optimization Problems, and Primal-dual Algorithms for Online Optimization.
Speakers:

  * Anupam Gupta, Carnegie Mellon University
  * Stefano Leonardi, Sapienza University of Rome
  * Seffi Naor, Technion

- Workshop "Research meets Industry" within the framework of the DFG priority programme "Algorithm Engineering" organized by Hannah Bast, Google Zürich, Switzerland, September 2008.

- Annual meeting of the DFG funded priority programme "Algorithm Engineering", organized by Ernst Althaus und Benjamin Doerr, October 2008.

28.14 Teaching Activities

Summer Semester 2007

Courses:

Optimization I (N. Garg, K. Elbassioni, and S. Canzar)

Machine Learning (H. Bast, K. Chang, S. Funke, and J. Giesen)

Winter Semester 2007/2008

Courses:

Algorithms and Datastructures (E. Althaus and B. Doerr)

Summer Semester 2008

Courses:

Tropical Geometry and Algebraic Statistics (O. Amini)

Internet Economics (G. Christodoulou, K. Elbassioni, and J. Mestre)

Optimization (E. Althaus and A. Karrenbauer)

Seminars:

Bio-inspired Computation (T. Friedrich and F. Neumann)

Winter Semester 2008/2009

Courses:

Randomized Algorithms (Selected Topics) (B. Doerr, A. Huber, and D. Johannsen)

Approximation Algorithms (S. Angelopoulos, N. Megow, and R. Raman)

Nonlinear Computational Geometry (M. Sagraloff and M. Hemmer)

Seminars:

Optimization of Submodular Functions and Applications (K. Elbassioni, J. Mestre, and R. Raman)

Diploma Theses

- Heiko Fisch, *Dominanzprobleme auf d-Trapezgraphen und Bandweite von Kettengraphen*, 2008
- Christian Klein-Heyl, *Improving Search Engines: Finding Infixes and Phrases*, 2007
- Patrick Pekczynksi, *Domain approximations for finite set constraint variables*, 2007

Bachelor Theses

- Tim Benke, *Semantic Overlay Networks for PLP Web Search*, 2007
- Marc Meier, *Complexity of search problems and Nash equilibra*, 2007
- Hung Son Pham, *Zeitplanung für saarländische Schulen mit ganzzahliger linearer Programmierung*, 2009
- Oliver Schönleben, *Six Degrees of Benchmarking Hypervolume Algorithms*, 2008
- Ching Hoo Tang, *Ant Colony Optimization using a Single Ant*, 2008
- Nima Zeini Jahromi, *Smoothed Analysis of Quicksort with Median-of-Three Pivot Rule*, 2008

Master Theses

- Irina Brudaru, *Heuristics for Average Diameter Approximation with external Memory Algorithms*, 2008
- Manuel Caroli, *Evaluation of a Generic Method for Analyzing Controlled-Pertubation Algorithms*, 2007
- Marjan Celikik, *Efficient Large-Scale Clustering of Spelling Variants*, 2007
- Alexandru Chitea, *Efficient Semantic Annotation of the English Wikipedia*, 2007
- Daniel Dumitriu, *Graph-based Conservative Surface Reconstruction*, 2008
- Pavel Emeliyanenko, *Visualization of Points and Segments of Real Algebraic Plane Curves*, 2007
- Alberto Escalante, *Privacy-Protecting Multi-Coupon Schemes with Stronger Protection against Splitting*, 2008
- Sebastian Limbach, *Towards the Exact Arrangement of Quadrics*, 2008
- Madhusudan Manjunath, *The Complexity of Minimizing absolute Gaussian curvature and computing all solutions for instances of the slab support vector machine problem*, 2008
- Gabriel Manolache, *Index-Based Snippet Generation*, 2008

- Momchil Rusinov, *Homomorphism Homogeneous Graphs*, 2008
- Daniel Schmitt, *WNETS - A Framework for Testing and Evaluating Algorithms and Models for Wireless Sensor Networks*, 2007

28.15 Dissertations, Habilitations, Offers, Awards

28.15.1 Dissertations

- Deepak Ajwani, *Traversing Large Graphs in Realistic Settings*, 2008
- Eric Berberich, *Robust and Efficient Software for Problems in 2.5-Dimensional Geometry Algorithms and Implementation*, 2008
- Markus Behle, *Binary Decision Diagrams and Integer Programming*, 2007
- Stefan Canzar, *Langrangian Relaxation Solving NP-hard Problems in Computational Biology via Combinatorial Optimization*, 2008
- Arno Eigenwillig, *Real Root isolation for Exact and Approximate Polynomials Using Descartes' Rule of Signs*, 2008
- Tobias Friedrich, *Use and Avoidance of Randomness*, 2007
- Tobias Gärtner, *Analytische Maschinen und Berechenbarkeit analytischer Funktionen*, 2008
- Edda Happ, *Analyses of Evolutionary Algorithms*, 2009
- Michael Hemmer, *Exact Computation of the Adjacency Graph of an Arrangment of Quadrics*, 2008
- Andreas Karrenbauer, *Enineering Combinatorial Optimization Algorithms to Improve the Lifetime of OLED Displays*, 2007
- Sören Laue, *Approximation Algorithms for Geometric Optimization Problems*, 2008
- Debapriyo Majumdar, *On Spetral Retrieval and Efficient Top-k Query Processing*, November 2007
- Domagoj Matijevic, *Geometric Optimization and Querying*, 2007
- Rouven Naujoks, *NP-hard Networking Problems*, 2008
- Hans Raj Tiwary, *Complexity of Some Polyhedral Enumeration Problems*, 2008
- Ingmar Weber, *Efficient Data Structures for and Applications of the CompleteSearch Engine*, November 2007

28.15.2 Habilitations

Completed:

- Ernst Althaus, 2008.
- Holger Bast, 2008.
- Lutz Kettner, 2008.
- Rob van Stee, *Combinatorial algorithms for packing and scheduling problems*, 2008.

28.15.3 Offers for Faculty Positions

- Ernst Althaus, Universität Mainz, W2-Professorship
- Omid Amini, Ecole Normale Supérieure, Paris, CNRS Researcher
- Hannah Bast, Universität Freiburg, W3-Professorship
- Hannah Bast, Universität Mainz, W3-Professorship
- Benjamin Doerr, RWTH Aachen, W2-Professorship
- Benjamin Doerr, TU Dortmund, W2-Professorship
- Stefan Funke, Universität Greifswald, W3-Professorship
- Joachim Giesen, Universität Bonn, W2-Professorship
- Joachim Giesen, Universität Jena, W3-Professorship
- Sathish Govindarajan (Assistant Professor, Indian Institute of Science, Bangalore)
- Domagoj Matijevic (Assistant Professor, Strossmayer University, Osijek, Croatia)
- Katarzyna Paluch (Assistant Professor, University of Wroclaw, Poland)

28.15.4 Awards

- Hannah Bast, Teaching Innovation Award, CS department of Saarland University, 2007.
- Hannah Bast and Stefan Funke, Heinz-Billing-Preis, 2007.
- Hannah Bast, Meyer-Struckmann Science Award, 2007.
- Hannah Bast and Stefan Funke, SaarLB Science Award 2008.
- Hannah Bast, Alcatel-Lucent Science Award Technical Communication 2008.
- Rolf Harren, Hans-Uhde-Preis zur Förderung der Wissenschaft, 2008.
- Kurt Mehlhorn, Honorary Doctorate Degree, Aarhus University, Denmark, 2008.

- Konstantinos Panagiotou, ETH Medal for PhD Thesis, 2008.
- The paper "On the Runtime Analysis of the 1-ANT ACO Algorithm" by B. Doerr, F. Neumann, D. Sudholt and C. Witt won the best paper award of the track "Ant Colony Optimization, Swarm Intelligence, and Artificial Immune Systems" at the Genetic and Evolutionary Computation Conference (GECCO) 2007.
- The paper "Adaptive Local Ratio" by Julian Mestre won the best student paper award at the ACM-SIAM Symposium on Discrete Algorithms (SODA) 2008.
- The paper "Crossover Can Provably be Useful in Evolutionary Computation" by B. Doerr, E. Happ and C. Klein won the best paper award of the track "Evolutionary Combinatorial Optimization" at the Genetic and Evolutionary Computation Conference (GECCO) 2008.
- The paper "Theoretical Analysis of Diversity Mechanisms for Global Exploration" by T. Friedrich, P. S. Oliveto, D. Sudholt and C. Witt won the best paper award of the track "Genetic Algorithms" at the Genetic and Evolutionary Computation Conference (GECCO) 2008.
- The paper "Approximating Minimum Multicuts by Evolutionary Multi-objective Algorithms" by F. Neumann and J. Reichel won the best paper award at the 10th International Conference on Parallel Problem Solving from Nature (PPSN) 2008.

28.16 Grants and Cooperations

- Ernst Althaus, DFG Grant "Simple and Fast Implementation of Exact Optimization Algorithms with SCIL" in the Priority Programme "Algorithm Engineering".
- Ernst Althaus and Kurt Mehlhorn, project in collaborative research center AVACS.
- Hannah Bast, DFG Grant "Efficient Search in Very Large Text Collections, Databases, and Ontologies" in the Priority Programme "Algorithm Engineering".
- Benjamin Doerr, DFG Grant "Algorithm Engineering for Randomized Rounding" in the Priority Programme "Algorithm Engineering".
- Benjamin Doerr and Frank Neumann, Collaborative Research Project on "Structural Characterization of Good Instances for Randomized Search Strategies" with Surender Baswana, Somenath Biswas, and Piyush P. Kurur, IIT Kanpur, 2008-2010.
- Amr Elmasry, Alexander von Humboldt Fellowship, 2007 - 2008
- Kurt Mehlhorn, GIF Grant with Uri Zwick, Tel Aviv University, till December 2008.
- Kurt Mehlhorn, GIF Grant with Dany Halperin, Tel Aviv University, since January 2009.

- Kurt Mehlhorn, Principal Investigator, Cluster of Excellence, Multimodal Computing and Interaction.
- Julián Mestre, Alexander von Humboldt Fellowship, 2007-2009.
- Frank Neumann, DFG Grant "Theoretical Foundations of Swarm Intelligence" with Carsten Witt, DTU Copenhagen, 2009-2011.
- Rob van Stee, DFG Grant (Eigene Stelle) "Approximation and Online Algorithms for Game Theory", 2007-2009.

28.17 Publications

Books

[1] R. Harren. *Mehrdimensionale Packungsprobleme - Approximation geometrischer Verallgemeinerungen klassischer Packungsprobleme*. VDM Verlag Dr. Müller, Saarbrücken, Germany, 2007.

[2] K. Mehlhorn and P. Sanders. *Algorithms and Data Structures: The Basic Toolbox*. Springer, Berlin, 2009.

Journal articles and book chapters

[1] D. J. Abraham, R. W. Irving, T. Kavitha, and K. Mehlhorn. Popular matchings. *SIAM Journal on Computing*, 37(4):1030–1045, 2007.

[2] N. Ahuja, A. Baltz, B. Doerr, A. Privetivy, and A. Srivastav. On the minimum load coloring problem. *Journal of Discrete Algorithms*, 5:533–545, 2007.

[3] E. Althaus and S. Canzar. A lagrangian relaxation approach for the multiple sequence alignment problem. *Journal of Combinatorial Optimization*, 16:127–154, 2008.

[4] S. Arya, T. Malamatos, and D. M. Mount. A simple entropy-based algorithm for planar point location. *ACM Transactions on Algorithms*, 3(2):17, 2007.

[5] S. Arya, T. Malamatos, D. M. Mount, and K. C. Wong. Optimal expected-case planar point location. *SIAM Journal on Computing*, 37(2):584–610, 2007.

[6] H. Bast, S. Funke, P. Sanders, and D. Schultes. Fast routing in road networks using transit nodes. *Science*, 316(5824):566, 2007.

[7] S. Baswana and S. Sen. A simple and linear time randomized algorithm for computing sparse spanners in weighted graphs. *Random Structures & Algorithms*, 30(4):532–563, 2007.

[8] N. Beldiceanu, I. Katriel, and S. Thiel. Filtering algorithms for the same and usedby constraints. *Archives of Control Sciences*, to appear, 2007. to appear.

[9] M. Bodirsky and M. Kutz. Determining the consistency of partial tree descriptions. *Artificial Intelligence*, 171(2/3):185–196, 2007.

[10] E. Boros, K. Elbassioni, V. Gurvich, and H. R. Tiwary. Characterization of the vertices and extreme directions of the negative cycle polyhedron and hardness of generating vertices of 0/1-polyhedra. *CoRR*, abs/0801.3790, 2008.

[11] D. Brockhoff, T. Friedrich, N. Hebbinghaus, C. Klein, F. Neumann, and E. Zitzler. On the effects of adding objectives to plateau functions. *IEEE Transactions on Evolutionary Computation*, 2009. To appear.

[12] P. Cameron, D. Johannsen, T. Prellberg, and P. Schweitzer. Counting defective parking functions. *Electronic Journal of Combinatorics*, 15(1):R92, 2008.

[13] I. Caragiannis, A. V. Fishkin, C. Kaklamanis, and E. Papaioannou. A tight bound for online colouring of disk graphs. *Theoretical Computer Science*, 384(2/3):152–160, 2007.

[14] T.-H. H. Chan. Metric embeddings with relaxed guarantees. *SIAM Journal on Computing*, 38:2303–2329, 2009.

[15] T.-H. H. Chan. Small hop-diameter sparse spanners for doubling metrics. *Discrete and Computational Geometry*, 41(1):28–44, 2009.

[16] G. Christodoulou. Price of anarchy. In M.-Y. Kao, ed., *Encyclopedia of Algorithms*, pp. 1–99. Springer US, New York, 2008.

[17] G. Christodoulou and E. Koutsoupias. Mechanism design for scheduling. *Bulletin of European Association for Theoretical Computer Science*, 2009.

[18] G. Christodoulou, E. Koutsoupias, and A. Kovács. Mechanism design for fractional scheduling on unrelated machines. *ACM Transactions on Algorithms*, 2009.

[19] G. Christodoulou, E. Koutsoupias, and A. Nanavati. Coordination mechanisms. *Theoretical Computer Science*, 2009.

[20] G. Christodoulou, E. Koutsoupias, and A. Vidali. A lower bound for scheduling mechanisms. *Algorithmica*, 2009.

[21] M. Chrobak, L. Gasieniec, and D. R. Kowalski. The wake-up problem in multihop radio networks. *SIAM Journal on Computing*, 36(5):1453–1471, 2007.

[22] R. Cole and L. Kowalik. New linear-time algorithms for edge-coloring planar graphs. *Algorithmica*, 50(3):351–368, 2008. To appear.

[23] O. Cooley, N. Fountoulakis, D. Kühn, and D. Osthus. 3-uniform hypergraphs of bounded degree have linear ramsey numbers. *Journal of Combinatorial Theory B*, 98:484–505, 2008.

[24] O. Cooley, N. Fountoulakis, D. Kühn, and D. Osthus. Embeddings and ramsey numbers of sparse k-uniform hypergraphs. *Combinatorica*, xx, 2009. to appear.

[25] J. Cooper, B. Doerr, T. Friedrich, and J. Spencer. Deterministic random walks on regular trees. *Electronic Notes in Discrete Mathematics*, 29:509–513, 2007.

[26] J. Cooper, B. Doerr, J. Spencer, and G. Tardos. Deterministic random walks on the integers. *European Journal of Combinatorics*, 28(8):2072–2090, 2007.

[27] B. Csaba. On the bollobas-eidrigde conjecture for bipartite graphs. *Combinatorics, Probability and Computing*, 16(1):1–31, 2007.

[28] F. Diedrich, R. Harren, K. Jansen, R. Thöle, and H. Thomas. Approximation algorithms for 3d orthogonal knapsack. *Journal of Science and Technology*, 23(5):749–762, 2008.

[29] B. Doerr. Matrix approximation and Tusnády's problem. *European Journal of Combinatorics*, 28(3):990–995, 2007.

[30] B. Doerr. Partial colorings of unimodular hypergraphs. *Electronic Notes in Discrete Mathematics*, 29:359–363, 2007.

[31] B. Doerr. Roundings respecting hard constraints. *Theory of Computing Systems*, 40(4):467–483, 2007.

[32] B. Doerr and T. Friedrich. Deterministic random walks on the two-dimensional grid. *Combinatorics, Probability and Computing*, 18:123–144, 2009.

[33] B. Doerr, T. Friedrich, C. Klein, and R. Osbild. Unbiased matrix rounding. *Electronic Notes in Discrete Mathematics*, 28:41–46, 2007.

[34] B. Doerr, M. Gnewuch, P. Kritzer, and F. Pillichshammer. Component-by-component construction of low-discrepancy point sets of small size. *Monte Carlo Methods and Applications*, 14:129–149, 2008.

[35] B. Doerr, N. Hebbinghaus, and F. Neumann. Speeding up evolutionary algorithms through asymmetric mutation operators. *Evolutionary Computation*, 15(4):401–410, 2007.

[36] F. Eisenbrand, S. Funke, A. Karrenbauer, and D. Matijevic. Energy-aware stage illumination. *International Journal of Computational Geometry and Applications*, 18(1/2):107–129, 2008.

[37] F. Eisenbrand, S. Funke, A. Karrenbauer, J. Reichel, and E. Schömer. Packing a truck - now with a twist! *International Journal of Computational Geometry & Applications*, 17(5):505–527, 2007.

[38] F. Eisenbrand, A. Karrenbauer, M. Skutella, and C. Xu. Multiline addressing by network flow. *Algorithmica*, 53(4):583–596, 2009.

[39] M. El Kahoui and S. Rakrak. Structure of Groebner bases with respect to block orders. *Mathematics of Computation*, 76(260):2181–2187, 2007.

[40] K. Elbassioni. A note on systems with max-min and max-product constraints. *Fuzzy Sets and Systems*, 159(17):2272–2277, 2008.

[41] K. Elbassioni. On the complexity of monotone dualization and generating minimal hypergraph transversals. *Discrete Applied Mathematics*, 156(11):2109–2123, 2008.

[42] K. Elbassioni. Algorithms for dualization over products of partially ordered sets. *SIAM J. Discrete Math*, 23(1):487–510, 2009.

[43] K. Elbassioni, A. Elmasry, and I. Kamel. Indexing schemes for multi-dimensional moving objects. In *Encyclopedia of GIS*, pp. 523–529. Springer, Berlin, 2008.

[44] K. Elbassioni and H. R. Tiwary. On computing the vertex centroid of a polyhedron. *CoRR*, abs/0806.3456, 2008.

[45] A. Elmasry and A. Hammad. Inversion-sensitive sorting algorithms in practice. *ACM Journal of Experimental Algorithmics*, 13(1):1–18, 2008.

[46] A. Elmasry, C. Jensen, and J. Katajainen. Multipartite priority queues. *ACM Transactions on Algorithms*, 5(1):1–19, 2008.

[47] A. Elmasry, C. Jensen, and J. Katajainen. Two new methods for constructing double-ended priority queues from priority queues. *Computing*, 83(4):193–204, 2008.

[48] L. Epstein, A. Levin, and R. van Stee. Online unit clustering: Variations on a theme. *Theoretical Computer Science*, 407(1-3):85–96, 2008.

[49] L. Epstein, A. Levin, and R. van Stee. Two-dimensional packing with conflicts. *Acta Informatica*, 45(3):155–175, 2008.

[50] L. Epstein and R. van Stee. On the online unit clustering problem. *ACM Transactions on Algorithms*, 2009. To appear.

[51] L. Epstein, R. van Stee, and T. Tamir. Paging with request sets. *Theory of Computing Systems*, 44(1):67–81, 2009.

[52] N. Fountoulakis. Percolation on sparse random graphs with given degree sequence. *Internet Mathematics*, xx, 2009. to appear.

[53] N. Fountoulakis, D. Kühn, and D. Osthus. The order of the largest complete minor in a random graph. *Random Structures and Algorithms*, 33:127–141, 2008.

[54] N. Fountoulakis and B. Reed. Faster mixing and small bottlenecks. *Probability Theory and Related Fields*, 137:475–486, 2007.

[55] N. Fountoulakis and B. Reed. The evolution of the mixing rate of a simple random walk on the giant component of a random graph. *Random Structures and Algorithms*, 33:68–86, 2008.

[56] T. Friedrich, J. He, N. Hebbinghaus, F. Neumann, and C. Witt. Analyses of simple hybrid algorithms for the vertex cover problem. *Evolutionary Computation,*, 17(1), 2009.

[57] T. Friedrich, N. Hebbinghaus, and F. Neumann. Comparison of simple diversity mechanisms on plateau functions. *Theoretical Computer Science*, 2009. To appear.

[58] T. Friedrich and F. Neumann. When to use bit-wise neutrality. *Natural Computing*, 2009. To appear.

[59] G. Froyland, T. Koch, N. Megow, E. Duane, and H. Wren. Optimizing the landside operation of a container terminal. *OR Spectrum*, 30(1):53–75, 2008.

[60] S. Funke, A. Kesselman, F. Kuhn, Z. Lotker, and M. Segal. Improved approximation algorithms for connected sensor cover. *Wireless Networks*, 13(2):153–164, 2007.

[61] R. Gandhi and J. Mestre. Combinatorial algorithms for data migration to minimize average completion time. *Algorithmica*, X, 2007.

[62] J. Giesen. The combinatorial structure of polyhedral choice based conjoint analysis. In A. Gustafsson, A. Herrmann, and F. Huber, eds., *Conjoint Measurement: Methods and Applications*, ch. 13, pp. 259–271. Springer, Heidelberg, Germany, 4 edition, 2007.

[63] J. Giesen. Conjoint analysis for measuring the perceived quality in volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1664–1671, 2008.

[64] J. Giesen. Delaunay triangulation approximates anchor hull. *Computational Geometry - Theory and Applications*, 36:131–143, 2008.

[65] J. Giesen. The flow complex: A data structure for geometric modeling. *Computational Geometry - Theory and Applications*, 39:178–190, 2008.

[66] J. Giesen, D. Mitsche, and E. Schuberth. Collaborative ranking: an aggregation algorithm for individuals' preference estimation. In M.-Y. Kao and X.-Y. Li, eds., *Algorithmic Aspects in Information and Management, Third International Conference, AAIM 2007*, Portland, USA, 2007, vol. 4508, pp. 58–67. Springer.

[67] J. Giesen, E. Schuberth, K. Simon, P. Zolliker, and O. Zweifel. Image-dependent gamut mapping as optimization problem. *IEEE Transactions on Image Processing*, 16(10):2401–2410, 2007.

[68] C. Gotsman, K. Kaligosi, K. Mehlhorn, D. Michail, and E. Pyrga. Cycle bases of graphs and sampled manifolds. *Computer Aided Geometric Design*, 24(8/9):464, 2007. accepted for publication in Computer Aided Geometric Design journal.

[69] S. Govindarajan, M. C. Dietze, P. K. Agarwal, and J. S. Clark. A scalable algorithm for dispersing population. *Journal of Intellingent Information Systems*, 29(1):39–61, 2008.

[70] P. Hachenberger, L. Kettner, and K. Mehlhorn. Boolean operations on 3d selective Nef complexes: Data structure, algorithms, optimized implementation and experiments. *Computational Geometry: Theory and Applications*, 38(1-2):64–99, 2007.

[71] R. Harren. Approximation algorithms for orthogonal packing problems for hypercubes. *Theoretical Computer Science*, to appear, 2009.

[72] R. Harren and R. van Stee. Absolute approximation ratios for packing rectangles into bins. *Journal of Scheduling*, 2009. To appear.

[73] S. Hert, M. Hoffmann, L. Kettner, S. Pion, and M. Seel. An adaptable and extensible geometry kernel. *Computational Geometry: Theory and Applications*, 38(1-2):16–36, 2007.

[74] T. Kavitha and K. Mehlhorn. Algorithms to compute minimum cycle basis in directed graphs. *Theory of Computing Systems*, 40(4):485–505, 2007.

[75] L. Kettner, K. Mehlhorn, S. Pion, S. Schirra, and C. Yap. Classroom examples of robustness problems in geometric computations. *Computational Geometry: Theory and Applications*, 40(1):61–78, 2008.

[76] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, and V. Gurvich. Generating all vertices of a polyhedron is hard. *Discrete & Computational Geometry*, 39(1-3):174–190, 2008.

[77] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, V. Gurvich, and K. Makino. Generating cut conjunctions in graphs and related problems. *Algorithmica*, 51(3):239–263, 2008.

[78] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, V. Gurvich, G. Rudolf, and J. Zhao. On short paths interdiction problems: Total and node-wise limited interdiction. *Theory Comput. Syst.*, 43(2):204–233, 2008.

[79] L. Khachiyan, E. Boros, K. Elbassioni, and V. Gurvich. A global parallel algorithm for the hypergraph transversal problem. *Information Processing Letters*, 101(4):148–155, 2007.

[80] L. Khachiyan, E. Boros, K. Elbassioni, and V. Gurvich. On the dualization of hypergraphs with bounded edge-intersections and other related classes of hypergraphs. *Theoretical Computer Science*, 382(2):139–150, 2007.

[81] L. Khachiyan, E. Boros, K. Elbassioni, and V. Gurvich. Generating all minimal integral solutions to and-or systems of monotone inequalities: Conjunctions are simpler than disjunctions. *Discrete Applied Mathematics*, 156(11):2020–2034, 2008.

[82] L. Khachiyan, E. Boros, K. Elbassioni, and V. Gurvich. On enumerating minimal dicuts and strongly connected subgraphs. *Algorithmica*, 50(1):159–172, 2008.

[83] L. Khachiyan, E. Boros, K. Elbassioni, V. Gurvich, and K. Makino. Dual-bounded generating problems: Efficient and inefficient points for discrete probability distributions and sparse boxes for multidimensional data. *Theoretical Computer Science*, 379(3):361–376, 2007.

[84] L. Khachiyan, E. Boros, K. Elbassioni, V. Gurvich, and K. Makino. Enumerating disjunctions and conjunctions of paths and cuts in reliability theory. *Discrete Applied Mathematics*, 155(2):137–149, 2007.

[85] S. O. Krumke, A. Schwahn, R. van Stee, and S. Westphal. A monotone approximation algorithm for scheduling with precedence constraints. *Operations Research Letters*, 36(2):247–249, 2008.

[86] M. Kutz, K. Elbassioni, I. Katriel, and M. Mahajan. Simultaneous matchings: Hardness and approximation. *J. Comput. Syst. Sci.*, 74(5):884–897, 2008.

[87] J. Lehnert and P. Schweitzer. The co-word problem for the Higman-Thompson group is context-free. *Bulletin of the London Mathematical Society*, 39(2):235–241, 2007.

[88] T. Malamatos. Lower bounds for expected-case planar point location. *Computational Geometry Theory and Applications*, 39(2):91–103, 2008.

[89] K. Mehlhorn, T. Kavitha, and R. Hariharan. Faster deterministic and randomized algorithms for minimum cycle basis in directed graphs. *SIAM Journal of Computing*, 38(4):1430–1447, 2008.

[90] K. Mehlhorn and D. Michail. Minimum cycle bases: Faster and simpler. *ACM Transactions on Algorithms*, 2009. accepted for publication.

[91] D. Michail. Reducing rank-maximal to maximum weight matching. *Theoretical Computer Science*, 389(1-2):125–132, 2007.

[92] F. Neumann. Expected runtimes of evolutionary algorithms for the eulerian cycle problem. *Computers and Operations Research*, 35(9):2750–2759, 2008.

[93] F. Neumann, D. Sudholt, and C. Witt. Analysis of different MMAS ACO algorithms on unimodal functions and plateaus. *Swarm Intelligence*, 3(1):35–68, 2009.

[94] F. Neumann and C. Witt. Runtime analysis of a simple ant colony optimization algorithm. *Algorithmica*, 2009. To appear.

[95] K. Pruhs, R. van Stee, and P. Uthaisombut. Speed scaling of tasks with precedence constraints. *Theory of Computing Systems*, 43(1):67–80, 2008.

[96] E. Pyrga. Efficient models for timetable information in public transportation systems. *ACM Journal of Experimental Algorithmics*, 12:1–39, 2007.

[97] P. Schweitzer. Using the incompressibility method to obtain local lemma results for Ramsey-type problems. *Information Processing Letters*, 109(4):229–232, 2009.

[98] V. Sharma. Complexity of real root isolation using continued fractions. *Theoretical Computer Science*, 409:292–310, 2008.

[99] R. van Stee. Packet switching in single buffer. In M.-Y. Kao, ed., *Encyclopedia of Algorithms*, pp. 1–99. Springer, Berlin, 1 edition, 2008.

[100] R. van Stee. Paging. In M.-Y. Kao, ed., *Encyclopedia of Algorithms*, pp. 1–99. Springer, Berlin, 1 edition, 2008.

[101] C. Xu, A. Karrenbauer, K. M. Soh, and C. Codrea. Consecutive multiline addressing: A scheme for addressing pmoleds. *Journal of the Society for Information Display*, 16(2):211–219, 2008.

[102] E. Zotenko, J. Mestre, D. O'Leary, and T. Przytyzka. Essential complex biological modules explain the centrality-lethality rule. *PLoS Computational Biology*, x, 2008.

Conference articles

[1] D. Ajwani, K. Elbassioni, S. Govindarajan, and S. Ray. Conflict-free coloring for rectangle ranges using $n^{.382}$ colors. In P. B. Gibbons and C. Scheideler, eds., *SPAA 2007: Proceedings of the 19th Annual ACM Symposium on Parallel Algorithms and Architectures*, San Diego, California, 2007, pp. 181–187. ACM.

[2] D. Ajwani, K. M. Elbassioni, S. Govindarajan, and S. Ray. Conflict-free coloring for rectangle ranges using $\tilde{O}(n^{.382+\epsilon})$ colors. In *19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA 07)*, San Diego, CA, USA, 2007, Proceedings. ACM. To Appear.

[3] D. Ajwani and T. Friedrich. Average-case analysis of online topological ordering. In T. Tokuyama, ed., *Algorithms and Computation, 18th International Symposium, ISAAC 2007, Sendai, Japan, December 17-19, 2007. Proceedings*, Sendai, Japan, 2007, *LNCS 4835*, pp. 464–475. Springer.

[4] D. Ajwani, I. Malinger, U. Meyer, and S. Toledo. Characterizing the performance of flash memory storage devices and its impact on algorithm design. In C. McGeoch, ed., *Proc. 7th Intern. Workshop on Experimental Algorithms (WEA)*, Provincetown, USA, 2008, *LNCS 5038*, pp. 208–219. Springer.

[5] D. Ajwani, S. Ray, R. Seidel, and H. R. Tiwary. On computing the centroid of the vertices of an arrangement and related problems. In F. Dehne, J.-R. Sack, and N. Zeh, eds., *10th Workshop on Algorithms and Data Structures (WADS)*, Halifax, Canada, 2007, *LNCS 4619*, pp. 519–528. Springer.

[6] M. Albrecht, A. Karrenbauer, and C. Xu. A clipper-free algorithm for efficient hw-implementation of local dimming led-backlight. In *Proceedings of the 28th International Display Research Conference (IDRC)*, Orlando, Florida, USA, 2008, pp. 286–289. SID Society for Information Display.

[7] E. Althaus, T. Baumann, E. Schömer, and K. Werth. Trunk packing revisited. In *Experimental Algorithms, 6th International Workshop*, Rome, Italy, 2007, *LNCS 4525*, pp. 420–432. Springer.

[8] E. Althaus and S. Canzar. A lagrangian relaxation approach for the multiple sequence alignment problem. In A. Dress, Y. Xu, and B. Zhu, eds., *Combinatorial Optimization and Applications, First International Conference, COCOA 2007*, Xi'an, Shaanxi, China, 2007, *LNCS 4616*, pp. 267–278. Springer.

[9] E. Althaus and S. Canzar. Lasa: A tool for non-heuristic alignment of multiple sequences. In *Bioinformatics Research and Development (BIRD?08)*, Vienna, Austria, 2008, pp. 489–498. Springer.

[10] E. Althaus, S. Canzar, K. Elbassioni, A. Karrenbauer, and J. Mestre. Approximating the interval constrained coloring problem. In *11th Scandinavian Workshop on Algorithm Theory (SWAT)*, Gothenburg, Sweden, 2008, *LNCS 5124*, pp. 210–221. Springer.

[11] E. Althaus, S. Canzar, M. Emmett, A. Karrenbauer, A. Marshall, A. Meyer-Bäse, and H. Zhang. Computing h/d-exchange speeds of single residues from data of peptic fragments. In *23rd Annual ACM Symposium on Applied Computing (SAC)*, Fortaleza, Brazil, 2008, pp. 1273–1277. ACM.

[12] E. Althaus and D. Dumitriu. Fast and accurate bounds on linear programs. In *8th International Symposium on Experimental Algorithms (SEA 2009)*, Dortmund, Germany, 2009, Lecture Notes in Computer Science. Springer. Accepted for publication.

[13] E. Althaus and R. Naujoks. Reconstructing phylogenetic networks with one recombination. In C. C. McGeoch, ed., *Experimental Algorithms, 7th International Workshop, WEA 2008*, Massachusetts, USA, 2008, *LNCS 5038*, pp. 275–288. Springer.

[14] S. Angelopoulos. Improved bounds for the online steiner tree problem in graphs of bounded edge-asymmetry. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, New Orleans, Louisiana, USA, 2007, pp. 248–257. SIAM.

[15] S. Angelopoulos. A near-tight bound for the online steiner tree problem in graphs of bounded asymmetry. In *16th Annual European Symposium on Algorithms (ESA 2008)*, Karlsruhe, 2008, *LNCS 5193*, pp. 76–87. Springer.

[16] S. Angelopoulos, R. Dorrigiv, and A. López-Ortiz. On the separation and equivalence of paging strategies. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, New Orleans, Louisiana, USA, 2007, pp. 229–237. SIAM.

[17] S. Angelopoulos, R. Dorrigiv, and A. Lopez-Ortiz. List update with locality of reference. In *8th Latin American Symposium on Theoretical Informatics (LATIN 2008)*, Buzios, Brasil, 2008, *LNCS 4957*, pp. 399–410. Springer.

[18] S. Angelopoulos, A. Lopez-Ortiz, and A. Hamel. Optimal scheduling of contract algorithms with soft deadlines. In *23rd National Conference on Artificial Intelligence (AAAI 2008)*, Chicago, 2008, pp. 868–873. AAAI press.

[19] S. Angelopoulos and P. Schweitzer. Paging and list update under bijective analysis. In C. Mathieu, ed., *20th ACM-SIAM Symposium on Discrete Algorithms (SODA 2009)*, New York, 2009, pp. 1136–1145. ACM press.

[20] B. Aronov, T. Asano, and S. Funke. Optimal triangulation with steiner points. In T. Tokuyama, ed., *Algorithms and Computation, 18th International Symposium, ISAAC 2007, Sendai, Japan, December 17-19, 2007. Proceedings*, Sendai, Japan, 2007, *LNCS 4835*, pp. 681–691. Springer.

[21] N. Bansal and H.-L. Chan. Weighted flow time does not admit o(1)-competitive algorithms. In C. Mathieu, ed., *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New York, USA, 2009, pp. 1238–1244. ACM Press.

[22] N. Bansal, H.-L. Chan, and K. Pruhs. Speed scaling with an arbitrary power function. In C. Mathieu, ed., *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New York, USA, 2009, pp. 693–701. ACM Press.

[23] R. Bar-Yehuda, G. Flysher, J. Mestre, and D. Rawitz. Approximation of partial capacitated vertex cover. In L. Arge, M. Hoffmann, and E. Welzl, eds., *15th Annual European Symposium on Algorithms*, Eilat, Israel, 2007, *LNCS 4698*, pp. 335–346. Springer.

[24] H. Bast, A. Chitea, F. Suchanek, and I. Weber. Ester: Efficient search in text, entities, and relations. In C. Clarke, N. Fuhr, and N. Kando, eds., *30th International Conference on Research and Development in Information Retrieval (SIGIR'07)*, Amsterdam, Netherlands, 2007, pp. 671–678. ACM.

[25] H. Bast, D. Majumdar, and I. Weber. Efficient interactive query expansion with completesearch. In M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, eds., *16th Conference on Information and Knowledge Management (CIKM'07)*, Lisboa, Portugal, 2007, pp. 857–860. ACM.

[26] H. Bast and I. Weber. The completesearch engine: Interactive, efficient, and towards ir & db integration. In G. Weikum, ed., *CIDR 2007, 3rd Biennial Conference on Innovative Data Systems Research*, Asilomar, CA, USA, 2007, pp. 88–95. VLDB Endowment.

[27] S. Baswana, S. Biswas, B. Doerr, T. Friedrich, P. P. Kurur, and F. Neumann. Computing single source shortest paths using single-objective fitness functions. In T. Jansen, I. Garibay, R. Wiegand, and A. S. Wu, eds., *Proceedings of the 10th International Workshop on Foundations of Genetic Algorithms (FOGA 2009)*, Orlando, USA, 2009. ACM. To appear.

[28] M. Behle. On threshold bdds and the optimal variable ordering problem. In A. Dress, Y. Xu, and B. Zhu, eds., *Combinatorial Optimization and Applications, First International Conference, COCOA 2007*, Xi?an, China, 2007, *LNCS 4616*, pp. 124–135. Springer.

[29] M. Behle, M. Jünger, and F. Liers. A primal branch-and-cut algorithm for the degree-constrained minimum spanning tree problem. In C. Demetrescu, ed., *Experimental Algorithms, 6th International Workshop, WEA 2007*, Rome, Italy, 2007, *LNCS 4525*, pp. 379–392. Springer.

[30] E. Berberich, M. Caroli, and N. Wolpert. Exact computation of arrangements of rotated conics. In *Proceedings of 23rd European Workshop on Computational Geometry*, Graz, Austria, 2007, pp. 231–234. Technische Universitaet Graz.

[31] E. Berberich, E. Fogel, D. Halperin, K. Mehlhorn, and R. Wein. Sweeping and maintaining two-dimensional arrangements on surfaces: A first step. In L. Arge, M. Hoffmann, and E. Welzl, eds., *Algorithms - ESA 2007, 15th Annual European Symposium*, Eilat, Israel, 2007, *LNCS 4698*, pp. 645–656. Springer.

[32] E. Berberich, E. Fogel, D. Halperin, and R. Wein. Sweeping and maintaining two-dimensional arrangements on surfaces. In *Proceedings of 23rd European Workshop on Computational Geometry*, Graz, Austria, 2007, pp. 223–226. Technische Universitaet Graz.

[33] E. Berberich and M. Kerber. Arrangements on surfaces of genus one: Tori and dupin cyclides. In S. Petitjean, ed., *24th European Workshop on Computational Geometry - Collection of Abstracts*, Nancy, France, 2008, pp. 209–212. An extended version of this article has appeared under the name "Exact Arrangements on Tori and Dupin Cyclides" in the Proceedings of the 2008 ACM Symposium on Solid and Physical Modeling, pp 59-66.

[34] E. Berberich and M. Kerber. Exact arrangements on tori and dupin cyclides. In E. Haines and M. McGuire, eds., *Proceedings of the 2008 ACM Symposium on Solid and Physical Modeling*, Stony Brook, USA, 2008, pp. 59–66. ACM.

[35] E. Berberich, M. Kerber, and M. Sagraloff. Exact geometric-topological analysis of algebraic surfaces. In M. Teillaud and E. Welzl, eds., *Proceedings of the 24th ACM Symposium on Computational Geometry*, College Park Maryland, USA, 2008, pp. 164–173. ACM.

[36] E. Berberich, M. Kerber, and M. Sagraloff. Geometric analysis of algebraic surfaces based on planar arrangements. In S. Petitjean, ed., *24th European Workshop on Computational Geometry - Collection of Abstracts*, Nancy, France, 2008, pp. 29–32. An extended version of this article has appeared under the name "Exact Geometric-Topological Analysis of Algebraic Surfaces" in the Proceedings of the 24th ACM Symposium on Computational Geometry, 2008, pp 164-173.

[37] E. Berberich and M. Meyerovitch. Computing envelopes of quadrics. In *Proceedings of 23rd European Workshop on Computational Geometry*, Graz, Austria, 2007, pp. 235–238. Technische Universitaet Graz.

[38] E. Berberich and M. Sagraloff. A generic and flexible framework for the geometrical and topological analysis of (algebraic) surfaces. In E. Haines and M. McGuire, eds., *Proceedings of the 2008 ACM Symposium on Solid and Physical Modeling*, Stony Brook, USA, 2008, pp. 171–182. ACM.

[39] U. Bhaskar, L. Fleischer, D. Hoy, and C.-C. Huang. Equilibria of atomic flow games are not unique. In C. Mathieu, ed., *20th Annual ACM-SIAM Symposium on Discrete Algorithms*, New York, U.S.A., 2009, pp. 748–757. Society for Industrial and Applied Mathematics.

[40] E. Boros, K. Borys, K. Elbassioni, V. Gurvich, K. Makino, and G. Rudolf. Generating minimal k-vertex connected spanning subgraphs. In G. Lin, ed., *Computing and Combinatorics, 13th Annual International Conference, COCOON 2007, Proceedings*, Banff, Canada, 2007, *LNCS 4598*, pp. 222–231. Springer.

[41] E. Boros, K. Elbassioni, V. Gurvich, K. Makino, and V. Oudalov. A complete characterization of nash-solvability of bimatrix games in terms of the exclusion of certain 2x2 subgames. In *Computer Science - Theory and Applications, Third International Computer Science Symposium in Russia, CSR 2008*, Moscow, Russia, 2008, vol. 5010, pp. 99–109. Springer.

[42] E. Boros, K. Elbassioni, and K. Makino. On berge multiplication for monotone boolean dualization. In *Automata, Languages and Programming, 35th International Colloquium, ICALP 2008*, Reykjavik, Iceland, 2008, *LNCS 5125*, pp. 48–59. Springer.

[43] K. Bringman and T. Friedrich. Approximating the least hypervolume contributor: NP-hard in general, but fast in practice. In *Proceedings of the 5th International Conference on Evolutionary Multi-Criterion Optimization (EMO 2009)*, Nantes, France, 2009. ACM.

[44] K. Bringmann and T. Friedrich. Approximating the volume of unions and intersections of high-dimensional geometric objects. In S.-H. Hong, H. Nagamochi, and T. Fukunaga, eds., *Proceedings of the 19th International Symposium on Algorithms and Computation (ISAAC 2008)*, Gold Coast, Australia, 2008, *LNCS 5369*, pp. 436–447. Springer.

[45] K. Bringmann and T. Friedrich. Don't be greedy when calculating hypervolume contributions. In T. Jansen, I. Garibay, W. R. Paul, and A. S. Wu, eds., *Proceedings of the 10th International Workshop on Foundations of Genetic Algorithms (FOGA 2009)*, Orlando, USA, 2009. ACM.

[46] D. Brockhoff, T. Friedrich, N. Hebbinghaus, C. Klein, F. Neumann, and E. Zitzler. Do additional objectives make a problem harder? In D. Thierens, ed., *Genetic and Evolutionary Computation Conference 2007*, London, UK, 2007, pp. 765–772. ACM.

[47] D. Brockhoff, T. Friedrich, and F. Neumann. Analyzing hypervolume indicator based algorithms. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, eds., *Parallel Problem Solving from Nature (PPSN X)*, Dortmund, Germany, 2008, *LNCS 5199*, pp. 651–660. Springer.

[48] M. Celikik and H. Bast. Fast error-tolerant search on very large texts. In D. Shin, ed., *The 24th Annual ACM Symposium on Applied Computing*, Honolulu, Hawaii, USA, 2009, *PROCEEDINGS OF THE 2009 ACM SYMPOSIUM ON APPLIED COMPUTING*, vol. 104092, pp. 1724–1731. ACM.

[49] H.-L. Chan, J. Edmonds, T.-W. Lam, L.-K. Lee, A. Marchetti-Spaccamela, and K. Pruhs. Nonclairvoyant speed scaling for flow and energy. In S. Albers and J.-Y. Marion, eds., *Proceedings of the 26th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Freiburg, Germany, 2009, pp. 255–264. IBFI.

[50] K. Chang. Multiple pass streaming algorithms for learning mixtures of distributions in r^d. In *Algorithmic Learning Theory, 18th International Conference, Proceedings*, Sendai, Japan, 2007, *Lecture Notes in Artifical intelligence*, vol. 4754, p. 16. Springer. This is an extended abstract. Full version has been submitted to a journal.

[51] K. Chang and A. Johnson. Online and offline selling in limit order markets. In *4th International Workshop on Internet and Network Economics (WINE 2008)*, Shanghai, China, 2008, *LNCS 5385*, pp. 41–52. Springer.

[52] G. Christodoulou, L. Gourves, and F. Pascual. Scheduling selfish tasks: About the performance of truthful algorithms. In G. Lin, ed., *Computing and Combinatorics, 13th Annual International Conference, COCOON 2007*, Banff, Canada, 2007, *LNCS 4598*, pp. 187–197. Springer.

[53] G. Christodoulou, E. Koutsoupias, and A. Kovács. Mechanism design for fractional scheduling on unrelated machines. In L. Arge, C. Cachin, T. Jurdzi?nski, and A. Tarlecki, eds., *Automata, Languages and Programming, 34th International Colloquium, ICALP 2007*, Wroc?aw, Poland, 2007, *LNCS 4596*, pp. 40–52. Springer. To appear at ICALP 07.

[54] G. Christodoulou, E. Koutsoupias, and A. Vidali. A lower bound for scheduling mechanisms. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Philadelphia, 2007, pp. 1163–1170. SIAM.

[55] G. Christodoulou, E. Koutsoupias, and A. Vidali. A characterization of 2-player mechanisms for scheduling. In D. Halperin and K. Mehlhorn, eds., *Algorithms - ESA 2008, 16th Annual European Symposium, Karlsruhe, Germany, September 15-17, 2008. Proceedings.*, Karlsruhe, Germany, 2008, pp. 297–307. Springer.

[56] G. Christodoulou, A. Kovács, and M. Schapira. Bayesian combinatorial auctions. In L. Aceto, I. Damgård, L. A. Goldberg, M. M. Halldórsson, A. Ingólfsdóttir, and I. Walukiewicz, eds., *Automata, Languages and Programming, 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7-11, 2008, Proceedings, Part I*, Reykjavik, Iceland, 2008, pp. 820–832. Springer.

[57] J. Cooper, B. Doerr, T. Friedrich, and J. Spencer. Deterministic random walks on regular trees. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008*, San Francisco, USA, 2008, pp. 766–772. ACM.

[58] F. Diedrich, R. Harren, K. Jansen, R. Thöle, and H. Thomas. Approximation algorithms for 3d orthogonal knapsack. In J.-Y. Cai, S. B. Cooper, and H. Zhu, eds., *Theory and Applications of Models of Computation*, Shanghai, China, 2007, *LNCS 4484*, pp. 34–45. Springer.

[59] F. Diedrich, B. Kehden, and F. Neumann. Multi-objective problems in terms of relational algebra. In R. Berghammer, B. Möller, and G. Struth, eds., *Relations and Kleene Algebra in Computer Science 2008*, Frauenwörth, Germany, 2008, *LNCS 4988*, pp. 84–98. Springer.

[60] F. Diedrich and F. Neumann. Using fast matrix multiplication in bio-inspired computation for complex optimization problems. In *IEEE Congress on Evolutionary Computation 2008*, Hong Kong, 2008, pp. 3828–3833. IEEE.

[61] B. Doerr, A. Eremeev, C. Horoba, F. Neumann, and M. Theile. Evolutionary algorithms and dynamic programming. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM. To appear.

[62] B. Doerr, T. Friedrich, M. Künnemann, and T. Sauerwald. Quasirandom rumor spreading: An experimental analysis. In I. Finocchi and J. Hershberger, eds., *Proceedings of the 10th Workshop on Algorithm Engineering and Experiments (ALENEX 2009)*, New York, USA, 2009, pp. 145–153. SIAM.

[63] B. Doerr, T. Friedrich, and T. Sauerwald. Quasirandom rumor spreading. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008*, San Francisco, USA, 2008, pp. 773–781. ACM.

[64] B. Doerr and M. Gnewuch. Construction of low-discrepancy point sets of small size by bracketing covers and dependent randomized rounding. In A. Keller, S. Heinrich, and H. Niederreiter, eds., *Monte Carlo and Quasi-Monte Carlo Methods 2006*, Ulm, Germany, 2008, pp. 299–312. Springer.

[65] B. Doerr, M. Gnewuch, N. Hebbinghaus, and F. Neumann. A rigorous view on neutrality. In *IEEE Congress on Evolutionary Computation 2007*, Singapore, 2007, pp. 2591–2597. IEEE.

[66] B. Doerr and E. Happ. Directed trees: A powerful representation for sorting and ordering problems. In *Proceedings of CEC 2008*, Hong Kong, 2008, pp. 3606–3613. IEEE.

[67] B. Doerr, E. Happ, and C. Klein. A tight bound for the (1+1)-ea on the single source shortest path problem. In *IEEE Congress on Evolutionary Computation 2007*, Singapore, 2007, pp. 1890–1895. IEEE.

[68] B. Doerr, E. Happ, and C. Klein. Crossover can provably be useful in evolutionary computation. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, Proceedings of the 10th annual conference on Genetic and evolutionary computation, pp. 539–546. ACM. Best paper award.

[69] B. Doerr, T. Jansen, and C. Klein. Comparing local and global mutations on bit-strings. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, Proceedings of the 10th annual conference on Genetic and evolutionary computation, pp. 929–936. ACM.

[70] B. Doerr and D. Johannsen. Adjacency list matchings — an ideal genotype for cycle covers. In D. Thierens, ed., *Genetic and Evolutionary Computation Conference (GECCO-2007)*, London, UK, 2007, pp. 1203–1210. ACM.

[71] B. Doerr and D. Johannsen. Refined runtime analysis of a basic ant colony optimization algorithm. In *IEEE Congress on Evolutionary Computation 2007*, Singapore, 2007, pp. 501–507. IEEE.

[72] B. Doerr, D. Johannsen, and C. H. Tang. How single ant aco systems optimize pseudo-boolean functions. In *Parallel Problem Solving from Nature ? PPSN X*, USA, Atlanta, 2008, *LNCS 5199*, pp. 378–388. Springer.

[73] B. Doerr, C. Klein, and T. Storch. Faster evolutionary algorithms by superior graph representation. In *First IEEE Symposium on Foundations of Computational Intelligence (FOCI-2007)*, Honolulu, USA, 2007, pp. 245–250. IEEE.

[74] B. Doerr, F. Neumann, D. Sudholt, and C. Witt. On the runtime analysis of the 1-ANT ACO algorithm. In D. Thierens, ed., *Genetic and Evolutionary Computation Conference 2007*, London, UK, 2007, pp. 33–40. ACM. Best paper award.

[75] B. Doerr and M. Theile. Improved analysis methods for crossover-based algorithms. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM. To appear.

[76] B. Doerr and M. Wahlström. Randomized rounding in the presence of a cardinality constraint. In I. Finocchi and J. Hershberger, eds., *Proceedings of the 10th Workshop on Algorithm Engineering and Experiments (ALENEX 2009)*, New York, USA, 2009, pp. 162–174. SIAM.

[77] D. Dumitriu, S. Funke, M. Kutz, and N. Milosavljevic. How much Geometry it takes to Reconstruct a 2-Manifold in r^3. In *10th Workshop on Algorithm Engineering and Experiments (ALENEX-2008)*, San Francisco, USA, 2008, pp. 65–74. SIAM.

[78] D. Dumitriu, S. Funke, M. Kutz, and N. Milosavljevic. On the locality of extracting a 2-manifold in r^3. In J. Gudmundsson, ed., *11th Scandinavian Workshop on Algorithm Theory (SWAT-2008)*, Göteborg, Sweden, 2008, *LNCS 5124*, pp. 270–281. Springer.

[79] D. Dumitriu, S. Funke, M. Kutz, and N. Milosavljevic. On the locality of extracting a 2-manifold in r^3. In S. Petitjean, ed., *Collection of abstracts of the 24th European Workshop on Computational Geometry*, Nancy, France, 2008, pp. 205–208.

[80] L. Dupont, M. Hemmer, S. Petitjean, and E. Schömer. Complete, exact and efficient implementation for computing the adjacency graph of an arrangement of quadrics. In L. Arge, M. Hoffmann, and E. Welzl, eds., *15th Annual European Symposium on Algorithms*, Eilat, Israel, 2007, *LNCS 4698*, pp. 633–644. Springer.

[81] M. Eigensatz, J. Giesen, and M. Manjunath. The solution path of the slab support vector machine. In P. Morin, ed., *The 20th Canadian Conference on Computational Geometry*, Mcgill University, Montreal, Canada, 2008, pp. 211–214. CCCG.

[82] A. Eigenwillig and M. Kerber. Exact and efficient 2d-arrangements of arbitrary algebraic curves. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA08)*, San Francisco, USA, 2008, pp. 122–131. ACM/SIAM.

[83] A. Eigenwillig, M. Kerber, and N. Wolpert. Fast and exact geometric analysis of real algebraic plane curves. In C. W. Brown, ed., *Proceedings of the 2007 International Symposium on Symbolic and Algebraic Computation*, Waterloo, Ontario, Canada, 2007, pp. 151–158. ACM.

[84] A. Eigenwillig, L. Kettner, and N. Wolpert. Snap rounding of Bézier curves. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry (SCG'07)*, Gyeongju, South Korea, 2007, pp. 158–167. ACM.

[85] F. Eisenbrand, A. Karrenbauer, and C. Xu. Algorithms for longer oled lifetime. In C. Demetrescu, ed., *WEA 2007*, Rome, Italy, 2007, *LNCS 4525*, pp. 338–351. Springer.

[86] K. Elbassioni, M. Hagen, and I. Rauf. Some fixed-parameter tractable classes of hypergraph duality and related problems. In *Parameterized and Exact Computation, Third International Workshop, IWPEC 2008*, Victoria, Canada, 2008, *LNCS 5018*, pp. 91–102. Springer.

[87] K. Elbassioni, R. Raman, S. Ray, and R. Sitters. On the approximability of the maximum feasible subsystem problem with 0/1-coefficients. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2009*, New York, NY, USA, 2009, pp. 1210–1219. SIAM.

[88] K. Elbassioni, R. Sitters, and Y. Zhang. A quasi-ptas for profit-maximizing pricing on line graphs. In L. Arge, M. Hoffmann, and E. Welzl, eds., *Algorithms - ESA 2007, 15th Annual European Symposium, Eilat, Israel, October 8-10, 2007, Proceedings*, Eilat, Israel, 2007, *LNCS 4698*, pp. 451–462. Springer.

[89] K. Elbassioni and H. R. Tiwary. On a cone covering problem. In *Proceedings of the 20th Annual Canadian Conference on Computational Geometry*, Montreal, Canada, 2008, pp. 171–174. CCCG.

[90] A. Elmasry. Pairing heaps with O(log log n) decrease cost. In *20th ACM-SIAM Symposium on Discrete Algorithms*, New York, USA, 2009, pp. 471–476. Society for Industrial and Applied Mathematics (SIAM).

[91] P. Emeliyanenko and M. Kerber. Visualizing and exploring planar algebraic arrangements – a web application. In M. Teillaud and E. Welzl, eds., *Proceedings of the 24th ACM Symposium on Computational Geometry*, College Park Maryland, USA, 2008, pp. 224–225. ACM.

[92] L. Epstein and R. van Stee. Approximation schemes for packing splittable items with cardinality constraints. In C. Kaklamanis and M. Skutella, eds., *Approximation and Online Algorithms, 5th International Workshop, WAOA 2007*, Eilat, Israel, 2008, *LNCS 4927*, pp. 232–245. Springer.

[93] L. Epstein and R. van Stee. Maximizing the minimum load for selfish agents. In E. S. Laber, C. Bornstein, L. T. Noguiera, and L. Faria, eds., *LATIN 2008: Theoretical Informatics, 8th Latin American Symposium*, Búzios, Brazil, 2008, *LNCS 4957*, pp. 264–275. Springer.

[94] L. Epstein and R. van Stee. On the online unit clustering problem. In C. Kaklamanis and M. Skutella, eds., *Approximation and Online Algorithms, 5th International Workshop, WAOA 2007*, Eilat, Israel, 2008, *LNCS 4927*, pp. 193–206. Springer.

[95] L. Epstein and R. van Stee. The price of anarchy on uniformly related machines revisited. In B. Monien and U.-P. Schroeder, eds., *Algorithmic Game Theory, First International Symposium, SAGT 2008*, Paderborn, Germany, 2008, *LNCS 4997*, pp. 46–57. Springer.

[96] T. Friedrich, J. He, N. Hebbinghaus, F. Neumann, and C. Witt. Approximating covering problems by randomized search heuristics using multi-objective models. In D. Thierens, ed., *Genetic and Evolutionary Computation Conference 2007*, London, UK, 2007, pp. 797–804. ACM.

[97] T. Friedrich, J. He, N. Hebbinghaus, F. Neumann, and C. Witt. On improving approximate solutions by evolutionary algorithms. In *IEEE Congress on Evolutionary Computation 2007*, Singapore, Singapore, 2007, pp. 2614–2621. IEEE.

[98] T. Friedrich and N. Hebbinghaus. Average update times for fully-dynamic all-pairs shortest paths. In S.-H. Hong, H. Nagamochi, and T. Fukunaga, eds., *Proceedings of the 19th International Symposium on Algorithms and Computation (ISAAC 2008)*, Gold Coast, Australia, 2008, *LNCS 5369*, pp. 693–704. Springer.

[99] T. Friedrich, N. Hebbinghaus, and F. Neumann. Plateaus can be harder in multi-objective optimization. In *IEEE Congress on Evolutionary Computation 2007*, Singapore, Singapore, 2007, pp. 2622–2629. IEEE.

[100] T. Friedrich, N. Hebbinghaus, and F. Neumann. Rigorous analyses of simple diversity mechanisms. In D. Thierens, ed., *Genetic and Evolutionary Computation Conference 2007*, London, UK, 2007, pp. 1219–1225. ACM. nominated for best paper award.

[101] T. Friedrich, C. Horoba, and F. Neumann. Runtime analyses for using fairness in evolutionary multi-objective optimization. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, eds., *Parallel Problem Solving from Nature (PPSN X)*, Dortmund, Germany, 2008, *LNCS 5199*, pp. 671–680. Springer.

[102] T. Friedrich, C. Horoba, and F. Neumann. Multiplicative approximations and the hypervolume indicator. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM. To appear.

[103] T. Friedrich and F. Neumann. When to use bit-wise neutrality. In *IEEE Congress on Evolutionary Computation 2008*, Hong Kong, 2008, pp. 997–1003. IEEE.

[104] T. Friedrich, P. Oliveto, D. Sudholt, and C. Witt. Theoretical analysis of diversity mechanisms for global exploration. In M. Keijzer, ed., *Proceedings of the 10th annual conference on Genetic and evolutionary computation (GECCO 2008)*, Atlanta, GA, USAA, 2008, pp. 945–952. ACM.

[105] T. Friedrich and T. Sauerwald. Near-perfect load balancing by randomized rounding. In *Proceedings of the 41st ACM Symposium on Theory of Computing (STOC 2009)*, Bethesda, Maryland, USA, 2009. ACM.

[106] S. Funke, S. Laue, and R. Naujoks. Minimum energy broadcast with few senders. In J. Aspnes, C. Scheideler, A. Arora, and S. Madden, eds., *Distributed Computing in Sensor Systems, Third IEEE International Conference, DCOSS 2007*, Santa Fe, USA, 2007, *LNCS 4549*, pp. 404–416. Springer.

[107] S. Funke, S. Laue, R. Naujoks, and Z. Lotker. Power assignment problems in wireless communication: Covering points by disks, reaching few receivers quickly, and energy-efficient travelling salesman tours. In S. E. Nikoletseas, B. S. Chlebus, D. B. Johnson, and B. Krishnamachari, eds., *Distributed Computing in Sensor Systems, 4th IEEE International Conference, DCOSS 2008*, Santorini Island, Greece, 2008, *LNCS 5067*, pp. 282–295. Springer.

[108] S. Funke and N. Milosavljevic. Guaranteed-delivery geographic routing under uncertain node locations. In *IEEE INFOCOM 2007, 26th IEEE International Conference on Computer Communications*, Anchorage, USA, 2007, pp. 1244–1252. IEEE.

[109] S. Funke and I. Rauf. Information brokerage via location-free double rulings. In E. Kranakis and J. Opatrny, eds., *Ad-Hoc, Mobile, and Wireless Networks, 6th International Conference, ADHOC-NOW 2007, Morelia, Mexico, September 24-26, 2007, Proceedings*, Morelia, Mexico, 2007, *LNCS 4686*, pp. 87–100. Springer.

[110] A. Gidenstam and M. Papatriantafilou. Lfthreads: A lock-free thread library. In E. Tovar, P. Tsigas, and H. Fouchal, eds., *11th International Conference On Principles Of Distributed Systems (OPODIS)*, Guadeloupe, French West Indies, France, 2007, *LNCS 4878*, pp. 217–231. Springer.

[111] J. Giesen. Medial axis approximation from inner voronoi balls: A demo of the mesecina tool. In *23d Annual ACM Symposium on Computational Geometry*, Gyeongju, South-Korea, 2007, pp. 123–124. ACM.

[112] J. Giesen, E. Schuberth, K. Simon, P. Zolliker, and O. Zweifel. Image-dependent gamut mapping as optimization problem. *IEEE Transactions on Image Processing*, 16(10):2401–2410, 2007.

[113] E. Happ, D. Johannsen, C. Klein, and F. Neumann. Rigorous analyses of fitness-proportional selection for optimizing linear functions. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, pp. 953–960. ACM.

[114] R. Harren and R. van Stee. Packing rectangles into 2 opt bins using rotations. In J. Gudmundsson, ed., *11th Scandinavian Workshop on Algorithm Theory*, Gothenburg, Sweden, 2008, *LNCS 5124*, pp. 306–318. Springer.

[115] M. Hemmer and D. Hülse. Generic implementation of a modular gcd over algebraic extension fields. In *25th European Workshop on Computational Geometry*, Brussels, Belgium, 2009, pp. 321–324. Université Libre de Bruxelles.

[116] M. Hemmer, S. Limbach, and E. Schömer. Continued work on the computation of an exact arrangement of quadrics. In *25th European Workshop on Computational Geometry*, Brussels, Belgium, 2009, pp. 313–316. Université Libre de Bruxelles.

[117] C. Horoba and F. Neumann. Benefits and drawbacks for the use of epsilon-dominance in evolutionary multi-objective optimization. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, pp. 641–680. ACM Press.

[118] C. Horoba and F. Neumann. Additive approximations of pareto-optimal sets by evolutionary multi-objective algorithms. In *Foundations of Genetic Algorithms 2009*, Orlando, USA, 2009. ACM. To appear.

[119] C.-C. Huang, T. Kavitha, D. Michail, and M. Nasre. Bounded unpopularity matchings. In *11th Scandinavian Workshop on Algorithm Theory (SWAT)*, Gothenburg, Sweden, 2008, pp. 127–137. Springer.

[120] T. Kavitha. On a special co-cycle basis of graphs. In *11th Scandinavian Workshop on Algorithm Theory (SWAT)*, Gothenburg, Sweden, 2008, pp. 343–354. Springer.

[121] S. Khuller, A. Malekian, and J. Mestre. To fill or not to fill: The gas station problem. In L. Arge, M. Hoffmann, and E. Welzl, eds., *15th Annual European Symposium on Algorithms*, Eilat, Israel, 2007, *LNCS 4698*, pp. 534–545. Springer.

[122] S. Khuller and J. Mestre. An optimal incremental algorithm for minimizing lateness with rejection. In *16th Annual European Symposium on Algorithms (ESA)*, Karlsruhe, Germany, 2008, *LNCS 5193*, pp. 601–610. Springer.

[123] R. Klein and M. Kutz. Computing geometric minimum-dilation graphs is np-hard. In M. Kaufmann and D. Wagner, eds., *Graph Drawing, 14th International Symposium, GD 2006*, Karlsruhe, Germany, 2007, *LNCS 4372*, pp. 196–207. Springer.

[124] S. Kratsch. Polynomial kernelizations for MIN F+Pi1 and MAX NP. In S. Albers and J.-Y. Marion, eds., *26th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Freiburg, Germany, 2009. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

[125] S. Kratsch and F. Neumann. Fixed-parameter evolutionary algorithms and the vertex cover problem. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM.

[126] J. Kroeske, A. Ghandar, Z. Michalewicz, and F. Neumann. Learning fuzzy rules with evolutionary algorithms - an analytic approach. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, eds., *Parallel Problem Solving from Nature (PPSN X)*, Dortmund, Germany, 2008, *LNCS 5199*, pp. 1051–1060. Springer.

[127] E. Krohn, K. Elbassioni, D. Matijevic, J. Mestre, and D. Severdija. Improved approximation algorithms for 1.5D terrain guarding. In *26th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Freiburg - Germany, 2009. Internationales Begegnungs und Forschungszentrum f"ur Informatik (IBFI).

[128] M. Kutz and P. Schweitzer. ScrewBox: a randomized certifying graph non-isomorphism algorithm. In D. Applegate and G. Brodal, eds., *9th Workshop on Algorithm Engineering and Experiments (ALENEX'07)*, New Orleans, USA, 2007, pp. 150–157. SIAM.

[129] K. Mehlhorn. Matchings in graphs variations of the problem. In A. Dress, Y. Xu, and B. Zhu, eds., *Combinatorial Optimization and Applications, First International Conference, COCOA 2007*, Xi?an, China, 2007, *LNCS 4616*, pp. 1–2. Springer.

[130] K. Mehlhorn. Minimum cycle bases in graphs algorithms and applications. In L. Kucera and A. Kucera, eds., *Mathematical Foundations of Computer Science 2007, 32nd International Symposium, MFCS 2007*, C? esk? Krumlov, Czech Republic, 2007, *LNCS 4708*, pp. 13–14. Springer.

[131] J. Mestre. Adaptive local ratio. In *19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, San Francisco, USA, 2008, pp. 152–160. Society for Industrial and Applied Mathematics.

[132] J. Mestre. Lagrangian relaxation and partial cover (extended abstract). In *25th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Bordeaux, France, 2008, pp. 539–550. Internationales Begegnungs- und Forschungszentrum f"ur Informatik (IBFI), Schloss Dagstuhl, Germany.

[133] U. Meyer. On dynamic breadth-first search in external-memory. In S. Albers, P. Weil, and C. Rochange, eds., *Proc. 25th Annual Symposium on Theoretical Aspects (STACS)*, Bordeaux, France, 2008, *Dagstuhl Seminar Proceedings*, vol. 08001, pp. 551–560. IBFI.

[134] U. Meyer. On trade-offs in external-memory diameter approximation. In *Proc. 11th Scandinavian Workshop on Algorithm Theory (SWAT)*, Gothenburg, Sweden, 2008, *LNCS xxxx*. Springer.

[135] F. Neumann, P. S. Oliveto, and C. Witt. Theoretical analysis of fitness-proportional selection: Landscapes and efficiency. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009. ACM. To appear.

[136] F. Neumann and J. Reichel. Approximating minimum multicuts by evolutionary multi-objective algorithms. In G. Rudolph, T. Jansen, S. M. Lucas, C. Poloni, and N. Beume, eds., *Parallel Problem Solving from Nature (PPSN X)*, Dortmund, Germany, 2008, *LNCS 5199*, pp. 72–81. Springer. Best Paper Award.

[137] F. Neumann, J. Reichel, and M. Skutella. Computing minimum cuts by randomized search heuristics. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, pp. 779–786. ACM Press.

[138] F. Neumann, D. Sudholt, and C. Witt. Comparing variants of mmas aco algorithms on pseudo-boolean functions. In T. Stützle, M. Birattari, and H. H. Hoss, eds., *Engineering Stochastic Local Search Algorithms 2007*, Brussels, Belgium, 2007, *LNCS 4638*, pp. 61–75. Springer.

[139] F. Neumann, D. Sudholt, and C. Witt. Rigorous analyses for the combination of ant colony optimization and local search. In M. Dorigo, M. Birattari, C. Blum, M. Clerc, T. Stützle, and A. F. T. Winfield, eds., *International Conference on Ant Colony Optimization and Swarm Intelligence 2008*, Brussels, Belgium, 2008, *LNCS 5217*, pp. 132–143. Springer.

[140] F. Neumann and C. Witt. Ant colony optimization and the minimum spanning tree problem. In V. Maniezzo, R. Battiti, , and J.-P. Watson, eds., *International Conference on Learning and Intelligent Optimization 2007*, Trento, Italia, 2008, *LNCS 5313*, pp. 153–166. Springer.

[141] P. S. Oliveto, P. K. Lehre, and F. Neumann. Theoretical analysis of rank-based mutation - combining exploration and exploitation. In *IEEE Congress on Evolutionary Computation 2009*, Trondheim, Norway, 2009. IEEE. To appear.

[142] K. Panagiotou. Blocks in constrained random graphs with fixed average degree. In *21st International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC '09)*, Hagenberg, Austria, 2009. DMTCS (Discrete Mathematics and Theoretical Computer Science).

[143] K. Panagiotou and A. Steger. Maximal biconnected subgraphs of random planar graphs. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '09)*, January 4-6, 2009, 2009, pp. 432–440. ACM, SIAM.

[144] E. Pyrga and S. Ray. New existence proofs for epsilon-nets. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, College Park, MD, USA, 2008, pp. 199–207. ACM.

[145] S. Ray and N. H. Mustafa. Weak ϵ-nets have a basis of size $O(1/\epsilon \log 1/\epsilon)$ in any dimension. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry (SCG'07)*, Gyeongju, South Korea, 2007, pp. 239–244. ACM.

[146] N. Sherashidze, S. Vishwanathan, T. Petri, K. Mehlhorn, and K. Borgwardt. Efficient graphlet kernels for large graph comparison. In *12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Clearwater Beach, Florida USA, 2009, pp. 488–495. Society for Artificial Intelligence and Statistics.

[147] H. R. Tiwary and K. Elbassioni. On the complexity of checking self-duality of polytopes and its relations to vertex enumeration and graph isomorphism. In *Symposium on Computational Geometry 2008*, College Park, MD, USA,, 2008, pp. 192–198. ACM.

[148] M. Wahlström. A tighter bound for counting max-weight solutions to 2SAT instances. In M. Grohe and R. Niedermeier, eds., *3rd International Workshop on Parameterized and Exact Computation (IWPEC 2008)*, Victoria (BC), Canada, 2008, *LNCS 5018*, pp. 202–213. Springer.

[149] M. Wahlström. New plain-exponential time classes for graph homomorphism. In *Fourth International Computer Science Symposium in Russia, CSR 2009*, Novosibirsk, Russia, 2009. Springer. To appear.

[150] G. Weikum, H. Bast, G. Canright, D. Hales, C. Schindelhauer, and P. Triantafillou. Towards peer-to-peer web search. In *1st European Conference on Complex Systems, ECCS'05*, Paris, France, 2007. tba.

[151] C. Xu, A. Karrenbauer, K. M. Soh, and J. Wahl. A New Addressing Scheme for PM OLED Display. In J. Morreale, ed., *SID 2007 International Symposium Digest of Technical Papers*, Long Beach, USA, 2007, vol. XXXVIII, pp. 97–100. Society for Information Display.

Theses

[1] E. Berberich. *Robust and Efficient Software for Problems in 2.5-Dimensional Non-Linear Geometry - Algorithms and Implementations.* Phd thesis, Universität des Saarlandes, 2008.

[2] M. Caroli. Evaluation of a generic method for analyzing controlled-perturbation algorithms. Masters thesis, Universität des Saarlandes, 2007.

[3] D. Dumitriu. Graph-based conservative surface reconstruction. Masters thesis, Universität des Saarlandes, 2007.

[4] A. Eigenwillig. *Real Root Isolation for Exact and Approximate Polynomials Using Descartes? Rule of Signs.* Phd thesis, Universität des Saarlandes, Saarbrücken, 2008.

[5] P. Emeliyanenko. Visualization of points and segments of real algebraic plane curves. Masters thesis, Universität des Saarlandes, 2007.

[6] T. Friedrich. *Use and Avoidance of Randomness.* Phd thesis, Universität des Saarlandes, 2007.

[7] M. Hemmer. *Exact Computation of the Adjacency Graph of an Arrangement of Quadrics.* Phd thesis, Johannes Gutenberg-Universität Mainz, Saarstr. 21, D 55122 Mainz, 2008.

[8] A. Karrenbauer. *Engineering combinatorial optimization algorithms to improve the lifetime of OLED displays.* Phd thesis, Universität des Saarlandes, 2007.

[9] A. Kovács. *Fast Algorithms for Two Scheduling Problems.* Phd thesis, Universität des Saarlandes, 2007.

[10] S. Limbach. Continued work on the computation of an exact arrangement of quadrics. Masters thesis, Universität des Saarlandes, Saarbrücken, Germany, 2008.

[11] D. Matijevic. *Geometric Optimization and Querying - Exact and Approximate.* Phd thesis, Universität des Saarlandes, 2007.

[12] R. Naujoks. *NP-hard Networking Problems - Exact and Approximate Algorithms.* Phd thesis, Universität des Saarlandes, 2008.

[13] R. van Stee. *Combinatorial algorithms for packing and scheduling problems.* Habilitation thesis, Universität Karlsruhe, 2008.

[14] I. Weber. *Efficient Index Structures for and Applications of the CompleteSearch Engine.* Phd thesis, Universität des Saarlandes, 2007.

29 The Computational Biology and Applied Algorithmics Group (D3)

29.1 Personnel

Director
Prof. Dr. Dr. Thomas Lengauer

System Administration
Dr. Joachim Büch

Researchers
Dr. Mario Albrecht
Dr. Iris Antes (–August 2008)
Dr. Christoph Bock (full time until February 2009, part time since then)
Dr. Francisco Silva Domingues
Dr. Glenn Lawyer (March 2009–)
Dr. Gabriele Mayr
Dr. Ingolf Sommer
Dr. Hiroto Saigo (July 2008–)
Dr. Elena Zotenko (November 2007–)

PhD Students
Adrian Alexa
André Altmann
Yassen Assenov
Bastian Beggel (March 2009–)
Hagen Blankenburg (October 2007–)
Jasmina Bogojeska
Kasia Bozek (June 2007–)
Matthias Dietzen (November 2007–)
Dorothea Emig (June 2007–)
Lars Feuerbach (October 2007–)
Konstantin Halachev
Christoph Hartmann (–July 2008)
Lars Kunert (–July 2008)
Jochen Maydt (–July 2008)
Fidel Ramírez

Kirsten Roomp
Oliver Sander (–November 2008)
Andreas Schlicker
Sven-Eric Schelhorn (October 2007–)
Tobias Sing (–April 2007)
Andreas Steffen (–August 2007)
Priti Talwar (–November 2007)
Alexander Thielen
Laura Tolosi
Hongbo Zhu (–January 2009)

Secretary
Ruth Schneppen-Christmann

29.2 Visitors

In the time period from April 2007 to April 2009, the following researchers visited our group:

Dr. Robert Finn	03.05.07	Welcome Trust Sanger Institute, Hinxton, UK
Gregory Stephanopoulos	21.05.07	Massachusetts Institute of Technology, Cambridge, MA, USA
Steffi Benthin	12.06.07	University of Greifswald, Germany
Yungki Park	14.06.07	Saarland University, Saarbrücken, Germany
Susanne Kunkel	14.06.07	University of Heidelberg, Germany
Dr. Damian Fay	02.10.07	National University of Ireland, Galway, Ireland
Prof. Gyan Bhanot	24.10.07	BioMaps, Rutgers University, New Jersey, USA
Dr. Koji Tsuda	29.11.07	MPI for Biological Cybernetics, Tübingen, Germany
Prof. Rolf Backofen	21.01.08	University of Freiburg, Germany
Prof. Überla	22.01.08	Ruhr University, Bochum, Germany
Dave Richie	06.05.08	INRIA, Nancy, France
Prof. Frank Lammert	28.05.08	Saarland University Hospital, Homburg, Germany
Alberto de la Fuente	18.06.08–20.06.08	CRS4 Bioinformatica, Cagliari, Italy

Prof. Reinhard Jahn	08.07.08	MPI for Biophysical Chemistry, Göttingen, Germany
Prof. Silvio Tosatto	01.08.08–16.08.08	University of Padua, Italy
Emanuela Leonardi	01.08.08–16.08.08	University of Padua, Italy
Dr. Aleksejs Kontijevskis	12.08.08	Uppsala University, Uppsala, Sweden
Natalie Jäger	27.08.08	Johann-Wolfgang-Goethe University, Frankfurt, Germany
Yasuo Tabei	11.09.08	University of Tokyo, Japan
Dr. Hisashi Kashima	17.09.08	IBM Research Tokyo, Japan
Prof. Tatsuo Shioda	01.10.08	Research Institute of Microbial Diseases, Osaka University, Japan
Fabian Horn	25.11.08	University of Jena, Germany
Dr. Jérôme Waldispühl	11.03.09	Massachusetts Institute of Technology, Cambridge, MA, USA
Dr. Ville Mustonen	11.03.09	Institute of Theoretical Physics, University of Köln, Germany
Dr. Verena Wolf	12.03.09	Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
Inken Wohlers	06.04.09–07.04.09	Centrum Wiskunde & Informatica (CWI), Amsterdam, Netherlands

29.3 Department Organization

All members of the Department except the director, the secretary, and the system administrator have temporary positions. The Department has two research groups directed by scientists in the Department: Mario Albrecht directs the group *Molecular Networks in Medical Bioinformatics* and Christoph Bock directs the group *Computational Epigenetics*. (Since Christoph Bock has left for Harvard University/Broad Institute, the group is co-directed by him and Thomas Lengauer.) The two groups comprise five and three scientists, respectively, in addition to the group leaders. The remaining scientists report directly to the Director of the Department.

Doctoral students that are not close to finishing have regular meetings with the director (biweekly to bimonthly). In most cases, the supervision of doctoral students is assisted by scientists of the Department. Bachelor and masters students are generally advised by the scientists.

The Department meets regularly once a week (usually Thu 12:30-13:30) for a talk by one of the scientists or students (in rarer cases also by an external guest) and for discussions and announcements of general interest. Student thesis talks have become so frequent that they are often scheduled separately. The Department maintains a set of only internally accessible web

pages that collect information relevant for the Department such as central locally maintained databases, the status of student projects, seminar schedules and the like.

29.4 HIV Bioinformatics

Coordinator: Thomas Lengauer

By the end of 2007 about 33.2 million people were living with a human immunodeficiency virus type 1 (HIV-1) infection. Since first discovered, HIV is held responsible for about 25 million deaths worldwide. Despite 25 years of research neither a cure nor a vaccine for HIV is in sight. However, today's modern antiretroviral therapy has access to about 25 antiretroviral agents, which target different stages of the viral life cycle. Entry inhibitors and coreceptor antagonists aim at preventing entry of the virus into the host cell (see also: 29.4.2). The reverse transcriptase (RT), which transcribes viral RNA into DNA, is inhibited by two classes of RT inhibitors (RTIs). The viral enzyme integrase catalyzes the integration of the viral DNA in to the host genome, this step is suppressed by the newly approved class of integrase inhibitors. The drug class of protease inhibitors (PIs) inhibit the viral enzyme protease, which is involved in cleaving viral precursor polyproteins into functionally active units.

Our department has engaged in bioinformatical research on HIV for almost ten years. For many years bioinformatical support for analysis of viral resistance against classical drugs (RTIs and PIs) has been our main objective. Advances in this field are described in Section 29.4.1.

New drugs in development or on the marketplace target viral cell entry. The virus can evade such drugs by using a specific coreceptor for entering the host cell. Bioinformatic prediction methods for viral coreceptor usage are thus a topic of great interest in the medical community. Section 29.4.2 discusses our work on this topic. Section 29.4.3 summarizes research that we have performed on modeling viral evolution of HIV and coevolution of HIV (and its primate relative SIV) and the primate lineage. Such research eventually aims at unveiling the reasons why HIV is as pathogenic as it is (in contrast to some SIV variants). Finally, Section 29.4.4 summarizes additional work that has been performed in the Department on developing and applying new statistical methods to topics in the field of analyzing HIV resistance.

29.4.1 Resistance Analysis

Investigator: Andre Altmann

The large arsenal of antiretrovirals is the response to the virus' ability to escape from the medication by altering its genome. Since the RT lacks a proof-reading mechanism the mutation rate of HIV is extremely high: an estimated 10^{-4} to 10^{-6} substitutions per base pair per replication cycle. Due to the high mutation rate appearance of mutations conferring resistance to one or more antiretroviral drugs is very likely. These *drug resistance* mutations enable the virus to replicate despite the presence of the drug. Thus, the drug resistant variants will outgrow the wild-type virus in the population, and in the end render the drug useless. As a consequence modern antiretroviral therapy combines three or more drugs from different classes to ensure an effective and lasting suppression of viral replication. Despite the development of novel drugs and their improvement (in terms of adverse effects) the efficient

and optimized selection of compounds entering a regime is the key for ensuring a long lasting therapeutic success.

Prediction of Drug Resistance

It is possible to measure the drug resistance phenotype of the virus against a single drug in a laboratory assay (*phenotyping*). However, the procedure is time and labor intensive an requires high security levels. Thus, in recent years sequencing of genome regions targeted by antiretroviral drugs became standard of care *genotyping*. However, despite the time and cost advantage of genotyping over phenotyping, the derived information, namely the sequence, is harder to interpret than the result of the phenotyping experiment: the *resistance factor*, which is a single continuous value. To overcome this limitation expert boards derived sets of rules from clinical experience and literature to classify a virus (based on the genome sequence) to be either resistant or susceptible to a single drug. We approached this issue by predicting phenotype from genotype by means of statistical learning, i.e. GENO2PHENO uses support vector machines (SVMs) for predicting the resistance factor from the viral genotype [6]. GENO2PHENO has the appealing advantage of being independent from an expert board, and is therefore arguably more objective. However, training of the SVM models requires a critical amount of training data. These genotype-phenotype pairs are rarely available in sufficient amounts for novel compounds. To this end we investigated the use of semi-supervised statistical learning methods that facilitate construction of reliable models on the basis of few training samples and already available sequence data. Initial results [8] showed that for PIs and nucleoside RTIs semi-supervised methods can improve the performance over standard supervised methods. For non-nucleoside RTIs the approach failed. Cause for this failure is probably the simple resistance pattern of these drugs (i.e. one mutation is sufficient to cause complete resistance).

The newest addition to the geno2pheno web-services is an interpretation tool for resistance mutations against the new class of integrase inhibitors. GENO2PHENO[INTEGRASE] applies a set of expert rules that will be replaced by SVM models as soon as sufficient data are available.

Relevance of Predicted Phenotypes

In recent years (predicted) drug resistance phenotypes were criticized as a basis for inferring virological response to antiretroviral therapy (ART). Doubts were based on the assumption that mutations might have different effects *in vivo* than *in vitro*. Moreover, there are few mutations that are clinically relevant, but do not confer any resistance. In a recent study [4, 5], we compared predicted phenotypes with three expert-based interpretation approaches (HIVDB,ANRS, and REGA) for their ability to infer response to ART. In this study we used the commercially available predicted phenotype VIRCOTYPE that was trained on about 45,000 genotype-phenotype pairs per drug. The approaches were compared on a data set comprising about 6,300 treatment change episodes (TCEs). A TCE is defined as a treatment with associated viral genotype and binary outcome (success vs. failure). By simply summing up individual drug scores for computing a treatment score, expert-based and predicted phenotypes yielded comparable performance (measured via the area under the receiver

operator characteristics (ROC) curve (AUC)) ranging from 0.834 to 0.844. Application of the random forest statistical learning method for combining the individual drug scores to obtain a treatment score elevated the performance to a range from 0.895 to 0.921, where VIRCOTYPE performed best. Furthermore, this study revealed that a combination of predicted phenotypes and expert-based predictions (HIVDB+VIRCOTYPE) performs better than any combination of expert-based predictions alone, AUC of 0.928 compared to 0.912.

Optimization of Antiretroviral Therapy

As mentioned above, modern ART comprises three or more drugs from at least two different drug classes. However, interpretation approaches only predict resistance to single compounds. In clinical practice single drug scores are combined with simple schemes for deriving a treatment score. We previously introduced GENO2PHENO-THEO [1], which predicts response to ART on the basis of the viral genotype and the drugs in the regimen. The output of THEO is the estimated success probability of the regimen. Recently, we validated THEO on an independent data set comprising about 7,600 TCEs derived from the EuResist integrated database [2]. Performance of THEO was compared to the same three expert-based interpretation approaches. Figure 29.1 demonstrates the gain in sensitivity for all relevant specificities ($\leq 60\%$). Furthermore, we discovered that SVM models that were trained to predict response to a fixed drug combination could further improve the performance (see also: 29.4.4).

In the project EuResist funded by the European Union, in which our department has been a partner, the idea of THEO was brought to a European level. Aim of the EuResist project was the integration of multiple clinical databases to provide sufficient training data for a therapy response prediction tool. During the EuResist project three prediction engines were developed yielding comparable performance [9]. Moreover, various methods for combining the predictions provided by the single engine to a consensus outcome that could be displayed by the web-service were investigated [3]. Since July 2008 the EuResist combined prediction engine has been publicly available. Unlike in THEO covariates in addition to viral genotype and current regimen were explored within the EuResist project. The most promising additional information is the patient's therapeutic history [7] (see also: 29.4.4) and the viral load (VL; i.e. copies of HIV RNA per 1,000 ml blood) at treatment start. The therapeutic history can provide hints on pre-existing drug resistance mutations that are not visible in the current viral population persistent in the blood of the patient but are stored in tissues repositories. In an ongoing work we investigated the potential of predicted *viral fitness* as an additional predictor of response to ART. Here, viral fitness or replication capacity (RC) is the number of new viral particles generated within a fixed time in absence of any drug. We were able to predict RC using polynomial SVMs with satisfying performance ($\rho = 0.542$). However, the benefit of including predicted RC in models predicting change in VL or change in number of CD4 carrying cells (marker for the patient's immune status) was only moderate.

References

- [1] A. Altmann, N. Beerenwinkel, T. Sing, I. Savenkov, M. Däumer, R. Kaiser, S.-Y. Rhee, W. J. Fessel, R. W. Shafer, and T. Lengauer. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antiviral Therapy*, 12(2):169–178, 2007.

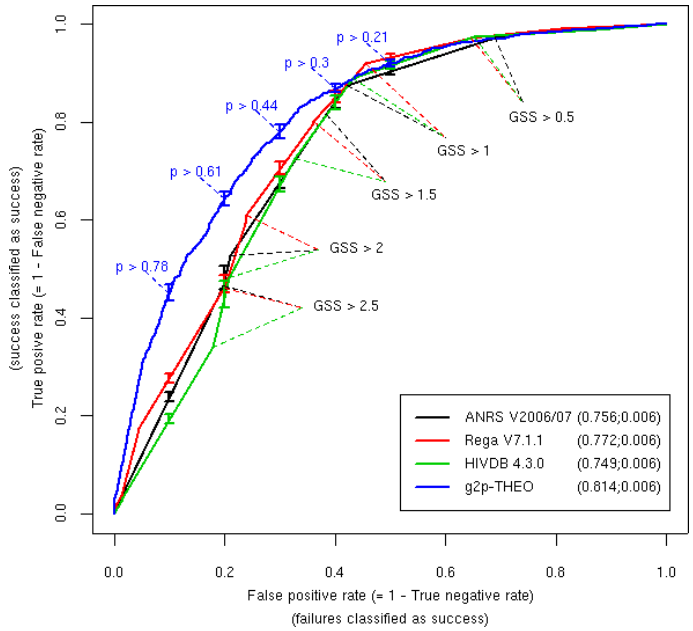


Figure 29.1: Receiver operating characteristic (ROC) curves for the data set used for validating geno2pheno-THEO (g2p-THEO). Every method is represented by a single ROC curve, namely HIVdb, ANRS, Rega, and g2p-THEO. Whiskers indicate the standard deviation of the true-positive rate (TPR) at a specific false-positive rate (FPR). The sum of single drug scores (GSS) and predicted success (p) cutoffs leading to specific TPR and FPR are indicated within the plot for expert-based approaches and g2p-THEO, respectively. For each method the AUC is given parenthetically in the box.

- [2] A. Altmann, M. Däumer, N. Beerenwinkel, Y. Peres, E. Schülter, J. Büch, S.-Y. Rhee, A. Sönnerborg, W. J. Fessel, R. W. Shafer, M. Zazzi, R. Kaiser, and T. Lengauer. Predicting the response to combination antiretroviral therapy: Retrospective validation of geno2pheno-THEO on a large clinical database. *The Journal of Infectious Diseases*, 199:999–1006, 2009.
- [3] A. Altmann, M. Rosen-Zvi, M. Prosperi, E. Aharoni, H. Neuvirth, A. Sönnerborg, E. Schülter, J. Büch, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi, and T. Lengauer. Comparison of classifier fusion methods for predicting response to anti HIV-1 therapy. *PLoS ONE*, 3(10):e3470, 2008.
- [4] A. Altmann, T. Sing, H. Vermeiren, B. Winters, E. Van Craenenbroeck, K. Van der Borght, S.-Y. Rhee, R. W. Shafer, E. Schülter, R. Kaiser, Y. Peres, A. Sonnerborg, W. J. Fessel, F. Incardona, M. Zazzi, L. Bacheler, H. Van Vlijmen, and T. Lengauer. Inferring virological response from genotype: with or without predicted phenotypes? In *Proceedings of the 16th International HIV Drug Resistance Workshop*, London, UK, 2007, vol. 12, p. S169. international medical.
- [5] A. Altmann, T. Sing, H. Vermeiren, B. Winters, E. Van Craenenbroeck, K. Van der Borght, S.-Y. Rhee, R. W. Shafer, E. Schülter, R. Kaiser, Y. Peres, A. Sönnerborg, W. J. Fessel, F. Incardona, M. Zazzi, L. Bacheler, H. Van Vlijmen, and T. Lengauer. Advantages of predicted phenotypes

and statistical learning models in inferring virological response to antiretroviral therapy from HIV genotype. *Antiviral Therapy*, 14(2):273–283, 2009.

[6] N. Beerenwinkel, M. Däumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter. Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Research*, 31(13):3850–3855, 2003.

- [7] F. Müller. Inferring virological response to antiretroviral combination therapy based on past treatment lines. Bachelor thesis, Universität des Saarlandes, 2008.
- [8] J. Perner. Semi-supervised learning for predicting anti-hiv drug resistance. Bachelor thesis, Universität des Saarlandes, 2008.
- [9] M. Rosen-Zvi, A. Altmann, M. Prosperi, E. Aharoni, H. Neuvirth, A. Sönnerborg, E. Schülter, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi, and T. Lengauer. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. In *Proceedings of the 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2008)*, Toronto, Canada, 2008, *Bioinformatics*, vol. 24, pp. i399–406. Oxford University.

29.4.2 Analysis of Coreceptor Usage

Investigators: Alexander Thielen and Kasia Bozek

HIV Cell Entry and Coreceptor Usage

HIV-1 cell entry is mediated by the cellular CD4-receptor and a second cellular membrane molecule, the so-called coreceptor, usually one of the chemokine receptors CCR5 and CXCR4. The binding of the viral envelope protein gp120 to the CD4-receptor induces a conformational switch in gp120 which reveals the coreceptor binding site. Subsequent binding of gp120 to the coreceptor leads to another conformational change of gp120 by which the screw-like viral gp41 protein is exposed. This protein then penetrates the cell membrane enabling the virus to fuse with the membrane of the cell and finally release the viral capsid into the cell.

Dependent on which coreceptor a virus uses it is called an R5-virus, if it only uses CCR5, an X4-virus, if it only uses CXCR4, or dual-tropic if it can infect cells via both coreceptors. Determination and monitoring of coreceptor usage is important for two reasons: First, the emergence of CXCR4-using variants has been correlated with progression to disease. More importantly, the first drug of a new class of antiretrovirals, the CCR5 antagnonist maraviroc, has been approved in the fall of 2007. Maraviroc targets CCR5, it is not effective against X4- or dual-tropic virus variants. Therefore testing coreceptor usage also known as tropism is mandatory before administration.

The routinely applied laboratory test in this context is the Monogram Trofile Tropism Assay, a phenotypic assay which is costly, time-consuming, and not easily accessible. A cheap and fast alternative is provided by genotypic approaches. These approaches are usually based on sequencing the third hypervariable (V3) loop of gp120 and predict coreceptor usage with interpretation tools. We have previously introduced the web-based tool geno2pheno[coreceptor] [4].

Previous Work on Coreceptor Usage in the Group

Coreceptor usage prediction has been consecutively improved over the last years [6]. While standard genotypic approaches infer viral tropism from sequence information alone, we

could demonstrate that providing explicit structural descriptors of the V3-loop to the prediction model improve predictions significantly [5]. In another work, we could show that incorporating immunological markers generally measured in clinics raises prediction quality substantially [7]. Models trained with these clinical parameters have also been implemented now in the geno2pheno[coreceptor] web service which is freely available on the Internet (http://www.geno2pheno.org). Since its start in 2004, the system has been used more than 100,000 times with around 80,000 predictions within the last two years.

Extending the Region of Interest – Prediction of HIV-1 Coreceptor Usage by Incorporating V2 Loop Sequence Variation

It is known that the V3 loop is the major determinant of coreceptor usage [2] and consequently this short region of about 35 residues is currently used in prediction tools. Nevertheless, several mutations outside V3, most notably within the bridging-sheet, have been shown to have an impact on coreceptor usage. We addressed the question to which extent non-V3 mutations can improve coreceptor usage prediction by analyzing the V2 loop of gp120, a region that has been repeatedly reported to have influence on coreceptor usage.

Clonal data, i.e. data where one specific viral strain is sequenced and measured in a phenotypic assay, were downloaded from the publicly available Los Alamos Sequence Database. Sequence analysis revealed not only several specific mutations but also some features like the number of N-glycosylation sites or the charge of the loop to be significantly associated with coreceptor usage. We incorporated these additional features into Support Vector Machines (SVMs) and tested against SVMs using only information of the V3-loop. In cross validation experiments we could show that the use of V2-sequence information leads to significant improvements over standard V3-models. The overall performance in terms of area under the receiver operator characteristics curve (AUC) increased from 0.914 to 0.933. Current genotypic approaches produce results at very high specificities (90%) , i.e. only a small fraction of the R5-viruses is incorrectly predicted as X4. When comparing our results at these levels we could observe that the sensitivity, the fraction of correctly classified X4-viruses, increases by about 4-5 percentage points when the V2-loop is used in the model.

We further validated our results on two *in vivo* datasets provided by our cooperation partners. The first dataset comprised therapy-naïve samples, i.e. samples from patients who were for the first time entering antiretroviral therapy in British Columbia whereas the second dataset contained isolates of therapy-experienced patients recently screened for maraviroc administration. Prediction models containing either sequence information of the V3-loop alone or in combination with the V2-loop were trained on the clonal dataset. On the therapy-naïve dataset consisting of 268 patient isolates, the standard V3-model was clearly outperformed by the prediction model using V2- and V3-loop data. In terms of AUC, the performance increased from 0.779 to 0.841. Similar results were found on the dataset of 64 therapy-experienced samples. There, AUC improved to 0.871 compared to 0.815 achieved with SVMs using only the V3-loop as input. These improvements were also observed at the interesting and clinically relevant specificity-level of 90%. When tested on therapy-naïve data, sensitivity of detecting X4-viruses increased from 54.2% to 62.8%. Even more profound differences were found on the therapy-experienced dataset where this number increased from 59.4% to 71.8% [8]. In summary, extending the genomic region on which we base prediction of coreceptor usage

beyond the V3 loop affords significantly more accurate predictions.

Inferring Viral Tropism from Genotype with Massively Parallel Sequencing: Qualitative and Quantitative Analysis

Previous works on coreceptor usage prediction have shown very good results when based on clonal data. However, the performance of all available prediction tools is dramatically decreased when tested on clinically derived data. The main difference between these two types of data is that the viral population *in vivo*, i.e., as extracted from the patient, is not a single clone but a swarm of genetically and phenotypically heterogeneous variants, a so-called *quasispecies*. When testing such a sample one usually generates a bulk-phenotype or a bulk-genotype which amounts to a genotypic or phenotypic consensus over the more frequent variants in the quasispecies. On the one hand such bulk sequencing data are much more difficult to interpret than clonal data, since the consensus is not an explicit representation of the occurring viral variants. On the other hand the bulk sequencing technology also incurs an inherent sensitivity problem: standard bulk-sequencing approaches can only detect genotypic minorities down to about 15-20% of the population. In comparison to standard bulk-sequencing, massively parallel sequencing based on the 454-technology can resolve the viral quasispecies to almost arbitrary detail. Thus, prediction on coreceptor usage based on 454 data should be significantly more accurate than on bulk data.

We performed a proof-of-concept study in which we combined massively parallel sequencing with our geno2pheno-tool [1]. Plasma samples from 55 antiretrovirally treated patients with tropism documented by the Monogram Trofile Tropism Assay were sequenced with standard bulk-sequencing approaches and their tropism was predicted with geno2pheno[coreceptor]. Specificity was at 90.9% while the sensitivity only reached 59.1%. From these isolates, 14 samples (7 R5, 7 X4) were selected for further analysis using 454-sequencing. Approximately 10,000 sequences per patient sample were generated and coreceptor usage inferred from them. For quantitative analysis of tropism distribution, the prediction score of each variant containing the V3 loop was plotted against its frequency within the viral population (see Fig. 29.2).

Minorities of sequences with high confidence in CXCR4-usage were found in all samples, irrespective of tropism. To obtain a summary value in order to compare with the phenotypic assay we adjusted the minority-level to the one proposed by the Trofile-assay and using the default false-positive-rate of geno2pheno[coreceptor] (10%). Thus we achieved concordance with the phenotypic output in all but one case. This suggests suggesting that the combination of this new sequencing technology with coreceptor usage prediction tools may be a fast and accurate alternative to phenotypic assays.

V3 Loop Sequence Space

The V3 loop of the HIV-1 envelope gene is a highly variable part of the viral genome. In order to provide a new perspective on the characteristics of CCR5- and CXCR4-tropic viral phenotypes we analyzed the distribution of the V3 loop sequences in sequence space. We used different similarity measures (Hamming, Blosum62 matrix), phylogenetic [3] and clustering methods in order to picture the distribution in sequence space of the V3 loops and their

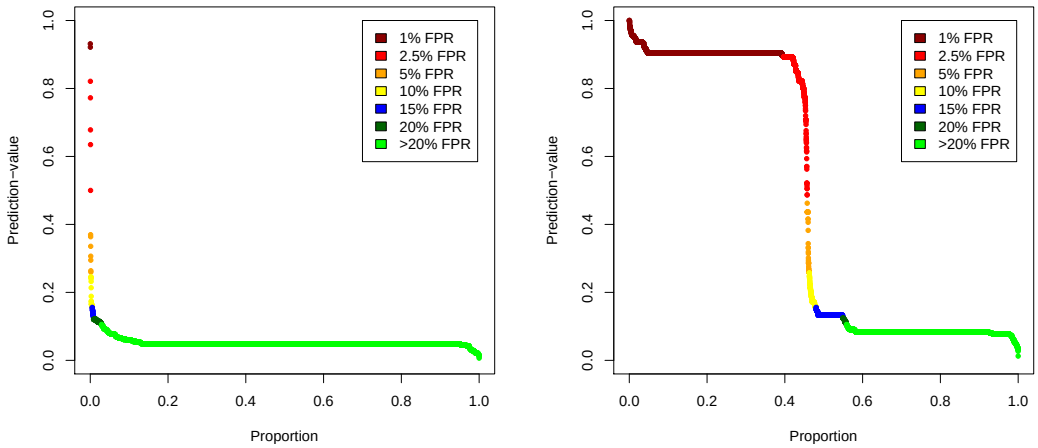


Figure 29.2: Quantitative analysis of coreceptor usage.

phenotype with respect to coreceptor usage.

References

- [1] M. Däumer, R. Kaiser, R. Klein, T. Lengauer, B. Thiele, and A. Thielen. Inferring viral tropism from genotype with massively parallel sequencing; qualitative and quantitative analysis. In *Antiviral Therapy*, London, UK, 2008, vol. 13, p. A101. international medical.

[2] R. A. Fouchier, M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J. Virol.*, 66:3183–3187, 1992.

[3] D. H. Huson. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14:68–73, 1998.

- [4] T. Lengauer, O. Sander, S. Sierra, A. Thielen, and R. Kaiser. Bioinformatics prediction of HIV coreceptor usage. *Nature Biotechnology*, 25(12):1407–1410, 2007.
- [5] O. Sander, T. Sing, I. Sommer, A. J. Low, P. K. Cheung, P. R. Harrigan, T. Lengauer, and F. S. Domingues. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLoS Computational Biology*, 3(3):555–564, 2007.
- [6] S. Sierra, R. Kaiser, A. Thielen, and T. Lengauer. Genotypic coreceptor analysis. *European Journal of Medical Research*, 12(9):453–462, 2007.
- [7] T. Sing, A. J. Low, N. Beerenwinkel, O. Sander, P. K. Cheung, F. S. Domingues, J. Büch, M. Däumer, R. Kaiser, T. Lengauer, and P. R. Harrigan. Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antiviral Therapy*, 12(7):1097–1106, 2007.
- [8] A. Thielen, P. R. Harrigan, A. J. Low, A. Moores, A. Altmann, K. Bozek, E. Heger, N. Sichtig, R. Kaiser, and T. Lengauer. Improved genotypic prediction of HIV-1 coreceptor usage by incorporating V2 loop sequence variation. In *Antiviral Therapy*, London, UK, 2008, vol. 13, p. A100. international medical.

29.4.3 Viral Evolution

Investigator: Kasia Bozek

Pathogens and hosts have mutual evolutionary effects on one another. In the host-pathogen interaction a constant interplay takes place between pathogen infectivity and host resistance, the ability of the host to clear an infection and the ability of the pathogen to evade or suppress host defenses. Example studies [3] show how after high initial pathogenicity due to the cross-species transmission event, both host and virus evolve towards more attenuated systems – the pathogen adapts to better survive in the new host, the host develops its ability to survive the infection. A highly pathogenic virus exerts a selection pressure on the host genome, favoring the fixation of resistance mutations in the population. On the other hand, because a longer life span of the host is beneficial for the pathogen increasing its chances of transmission, the virus will evolve towards a less pathogenic phenotype. Simian immunodeficiency viruses (SIVs) are primate lentiviruses that infect no fewer than 36 different nonhuman primate species in sub-Saharan Africa. Two of these viruses, SIVcpz from chimpanzees (Pan troglodytes) and SIVsmm from sooty mangabeys (Cercocebus atys), have crossed species barriers on multiple occasions and have generated human immunodeficiency virus (HIV) types 1 and 2 [2]. Primates naturally infected with SIV appear not to develop immunodeficiency or AIDS. In contrast, HIV infection of humans is nearly always characterized by progressive loss of $CD4^+$ T lymphocytes, chronic immune activation, and gradual destruction of immune functions. The basis for this difference in pathogenicity is not understood, but deciphering which viral and/or host factors are responsible for the nonpathogenic course of natural SIV infections could well prove useful in developing more-effective treatment or prevention strategies for HIV/AIDS.

We approach the question of HIV/SIV-host co-evolution on several different levels. First, as an example of a specific biological interaction between host and pathogen, we investigated how host immune recognition shapes the evolution of the viral V3 loop and how it results in different phenotypes 29.4.2. On a more general level, we compared evolutionary patterns of interacting host and viral genes in the primate lineage and among primate lentiviruses. Finally, we looked at reciprocal adaptive changes at the biochemical level – on molecule-for-molecule interactions between the host and pathogen.

TRIM5α Structure Evolution

TRIM5α is a restriction factor blocking the viral replication in a species-specific manner. For instance, restriction by rhesus monkey TRIM5α is efficient against HIV-1 but inefficient against SIVmac and undetectable against MLV [4]. The SPRY domain of TRIM5α is a major determinant for the virus restriction in primates. The role in anti-viral activity of V2 and V3 regions of the baboon TRIM5α has been examined experimentally in the lab of Prof. Tatsuo Shioda at the Department of Viral Infection, Research Institute for Microbial Diseases of Osaka University. We assisted the experimental work by providing a three-dimensional model of the investigated protein domain that allows to localize the variable regions on the protein structure. The model has been constructed by homology modeling based on recently published crystal structure of the PRYSPRY domain of mouse TRIM21 [1]. The variable region appears to form loops on the protein structure and its sequence composition might

influence the binding capacity of the protein to the viral capsid.

References

[1] A. H. Keeble, Z. Khan, A. Forster, and L. C. James. TRIM21 is an IgG receptor that is structurally, thermodynamically, and kinetically conserved. *Proc. Natl. Acad. Sci. U.S.A.*, 105:6045–6050, 2008.

[2] P. M. Sharp, G. M. Shaw, and B. H. Hahn. Simian immunodeficiency virus infection of chimpanzees. *J. Virol.*, 79:3891–3902, 2005.

[3] M. E. Woolhouse, J. P. Webster, E. Domingo, B. Charlesworth, and B. R. Levin. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.*, 32:569–577, 2002.

[4] L. M. Ylinen, Z. Keckesova, S. J. Wilson, S. Ranasinghe, and G. J. Towers. Differential restriction of human immunodeficiency virus type 2 and simian immunodeficiency virus SIVmac by TRIM5alpha alleles. *J. Virol.*, 79:11580–11587, 2005.

29.4.4 Statistical Methods

Investigators: Hiroto Saigo and Jasmina Bogojeska

In this section we describe additional statistical methods that we have applied to the problem of predicting resistance and therapy effectiveness in the context of HIV/AIDS. The first three subsections describe methods that deal with different types of bias in the clinical data sets. The fourth subsection predicts virus load by considering the interaction between mutations in a virus genotype [4].

Dealing with Bias in Clinical Data Sets

The data sets that contain information on the clinical outcome of therapies for many patients are biased in many ways. In what follows we give the main reasons that limit the objectiveness of the data. HIV data have been collected over two decades from different countries. Their therapy profiles differ regionally. Some therapies have been administered many times; others very few times if even at all. In order to enable therapy outcome prediction the data need to be labeled. This process requires several clinical measurements at regular time intervals. In the past, such measurements have not been carried out systematically, which results in a very sparse labeled data set. Furthermore, the viral sequences evolve over time and the treatment trends change with the introduction of new drugs.

The main goal of our research is bias-aware learning for HIV therapy screening. This includes developing new methods or adapting existing methods to the problem at hand and validating such methods. In the current stage our methods mainly follow the methodical paradigm of transfer learning.

Transfer Learning for HIV Therapy Screening

The main idea behind the transfer learning approach is to circumvent the deficiencies of the data sets by using additional information: for therapies with very few or no training examples we use the knowledge from similar therapies; to cope with the scarceness of the data set we use information from other sources (e.g. data from *in vitro* experiments).

We have developed a method [2] that trains a separate model for each particular therapy by using data from all available therapies from a clinical data set. The key point is that during learning we weight each data point properly. The weights are derived such that the distribution of the examples of all therapies is matched to the distribution of the therapy of interest. This approach enables predictions even for therapies with very few or no training examples and copes with the scarceness and uneven representation of therapies in the clinical databases. The predictions of the outcome of a therapy are based on the presence of a set of resistance-relevant mutations in the virus genotyped from the patient and his/hers historic treatment records. We additionally use prior knowledge on the similarity of different therapy combinations represented by suitable kernel functions. We provide two different kernel functions for this purpose. According to the first kernel the pair-wise similarities of two different therapy combinations are based on the set of common resistance-relevant mutations against their respective sets of drugs. The second kernel derives the similarities by comparing phenotypic information for the resistance against the drugs comprising the therapies of interest. In order to address the evolution of the viral genome in response to treatments over time we validate the method with time-consistent splits when choosing the training and the test sets. We calculate these splits by selecting the most recent samples as a test set and using the rest as a training set. In this way our models learn from data seen in the more distant past and their performance is measured on unseen data from the more recent past. Additionally we plan to do stratified cross-validation where the term stratified refers to even apportion of different combination therapies and class labels in the folds.

According to the experimental results obtained on the clinical data set from the EuResist project [1], 29.4.1 our method significantly improves the accuracy of predicting therapy outcome especially for therapies for which few training samples available.

Interaction Among Mutations and their Effects on Drug Resistance

A common approach to predicting the drug resistance provides regression models that treat mutations as variables, where each mutation is represented by a binary indicator variable representing its position in the wild type and the amino acid to which the wild type has mutated in this position. The resulting number of variables reaches into the hundreds. Conventional regression models, such as least squares regression, work well in this setting. However, such models do not consider interactions among mutations, in spite of the fact that such interactions are known to sometimes significantly influence resistance to drugs [3]. We developed a least squares model that can incorporate all the possible interaction terms between mutations [4]. In this case, the number of explanatory variables exceeds the number of response variables, so regularization is of importance. We employed L1 regularization, the state-of-the-art shrinkage and selection method in statistics, which gives us a sparse and interpretable solution. In experiments, our method worked particularly well for predicting the resistance of nucleotide reverse transcriptase inhibitors (NRTIs). Furthermore, it successfully recovered many mutation associations known in clinical literature.

References

- [1] A. Altmann, M. Rosen-Zvi, M. Prosperi, E. Aharoni, H. Neuvirth, E. Schülter, J. Büch, Y. Peres, I. Incardona, A. Sönnerborg, R. Kaiser, M. Zazzi, and T. Lengauer. The EuResist approach for

predicting response to anti HIV-1 therapy. In *Reviews in Antiviral Therapy*, Utrecht, Netherlands, 2008, vol. 2008, p. 107. Virology Education.

- [2] S. Bickel, J. Bogojeska, T. Lengauer, and T. Sheffer. Multi-task learning for HIV therapy screening. In A. McCallum and S. Roweis, eds., *Proceedings, Twenty-Fifth International Conference on Machine Learning*, Helsinki, Finland, 2008, pp. 56–63. Omnipress.

[3] A. K. N. Iverson, R. W. Shafer, K. Wehrly, M. A. Winters, M. J. I., B. Chesebro, and T. C. Merigan. Multidrug-resistant human immunodeficiency virus type 1 strains resulting from combination antiretroviral therapy. *Journal of Virology*, 70(2):1086–1090, 1996.

[4] H. Saigo, T. Uno, and K. Tsuda. Mining complex genotypic features for predicting hiv-1 drug resistance. *Bioinformatics*, 23(18):2455–2462, 2007.

29.5 Molecular Networks in Medical Bioinformatics

Coordinator: Mario Albrecht

Since its official establishment in 2006, the research group *Molecular Networks in Medical Bioinformatics* (`http://medbioinf.mpi-inf.mpg.de`) has been directed by Dr. Mario Albrecht in the Department of Computational Biology and Applied Algorithmics. The group also participates in the new MMCI Cluster of Excellence (`http://www.mmci.uni-saarland.de`) since 2008 and became a member of the local Center of Bioinformatics Saar in 2009. More information on the group itself can be found in Chapter 9 as well as online at `http://medbioinf.mpi-inf.mpg.de/group.php`. The group's external funding (see Chapter 29.14 for project details) includes two DFG-funded positions within the MMCI Cluster of Excellence, one DFG-funded position within the Clinical Research Group on hepatitis C, and one BMBF-funded position within the German NGFN initiative for environmental diseases, see Chapter 29.5.3 for medical details. The research group also contributes to the weekly departmental seminar series and maintains a public web site, an internal calendar and mailing list, and Wiki-based web pages as intranet for FAQs as well as further information about literature, research skills, and useful software.

The following three sections detail recent work within the reporting period of the last two years [9, 26, 21, 22, 1, 32, 30, 29, 15, 7, 5, 8, 6, 11, 20, 2]. All team members frequently attend national and international scientific meetings of bioinformaticians and biologists (e.g. GCB, ECCB, ISMB, ISCB, RECOMB, CSHL, DILS, ESF, etc.) and present their work in posters (listed on `http://medbioinf.mpi-inf.mpg.de/publications.php`) as well as in invited talks (∼30 talks within the reporting period, see detailed list in Chapter 29.11). Some PhD students also gave tutorials at international workshops about developed software tools (Cytoscape plugins DomainGraph and NetworkAnalyzer) and implemented web services and databases (FunSimMat, DASMIweb, BioMyn), see `http://medbioinf.mpi-inf.mpg.de/software.php` for details. In particular, both Cytoscape plugins and the online service FunSimMat have already attracted many biological users since their publication in 2008, while DASMIweb and BioMyn are new and available since 2009.

The group's research covers three overlapping research areas, namely, Network Analysis (Section 29.5.1), Modeling and Prediction (Section 29.5.2), and Disease Focus (Section 29.5.3). Different projects aim at computational approaches for analyzing molecular interaction networks as well as for modeling and predicting protein structure, function and interaction.

Particular interest lies in biological networks of regulatory and signaling processes, some of which are impaired in disease. The group develops and applies various bioinformatics methods, software tools, web services, and databases (`http://medbioinf.mpi-inf.mpg.de/software.php`) to gain comprehensive knowledge about the function and structure of interacting proteins that are as yet relatively uncharacterized, but of medical relevance. Currently, the disease focus is on autoinflammatory disorders and viral infections, for instance, on inflammatory bowel disease and hepatitis C. The computational findings support experimentally working partners in discovering promising candidate genes as well as in prioritizing and interpreting experiments targeted towards elucidating the molecular function of proteins and the cause of diseases.

Within the department D3, the research group collaborates with several colleagues on closely related bioinformatics topics concerning proteins and their interactions. In the course of the last years, it has also become apparent that the group's network-based research vision can be realized only in close collaborations with experimentally working biologists and medical researchers. Therefore, important cooperations have been established with scientists elsewhere in Germany, Europe and USA. A list of external partners involved with joint projects during the reporting period can be found on page 24 within the Overview chapter of the department D3. Projects together with other departments and research units on the Saarland University campus have not emerged yet, but there is a significant overlap of local research with regard to protein and network bioinformatics as well as to computer science (e.g. semantic web, network topology, information retrieval, statistical learning, graph algorithmics, large-scale data analysis and visualization). Within the Max Planck Society, interdisciplinary research and synergies might be further promoted by invitations to other sections and departments, organized meetings of young scientists and group leaders, and extra funds for innovative research and jointly supervised students. Since successful continuation and growth of research in network biology appears best feasible with cooperation partners nearby, the head of this group applies to other institutions in which systems biologists work and generate valuable experimental data for bioinformatics research.

29.5.1 Network Analysis

Investigators: Hagen Blankenburg, Fidel Ramírez, and Mario Albrecht

Exchange and Assessment of Molecular Interaction Data

In recent years, we could witness a substantial increase in the amount of publicly available molecular interaction data. This rapid accumulation of data is generated by several small- and large-scale experimental techniques, curation of the scientific literature, and numerous computational prediction methods. This diversity renders it more and more difficult for researchers to keep track of all the available information that is scattered across a multitude of online repositories. In the context of the EC-funded BioSapiens Network of Excellence [11], we have developed the Distributed Annotation System for Molecular Interactions (DASMI, `http://www.dasmi.de`), a new framework for exchanging, annotating, and assessing molecular interaction data [2]. The fundamental idea of DASMI is to avoid the unification of interaction data into a central repository by leaving the data with the original providers and retrieving and integrating them online on request. In contrast to other integration frameworks, our

approach eliminates the issue of centralized data maintenance and ensures that the accessed data are always up to date.

DASMI is based on the Distributed Annotation System (DAS) [4] and consists of a data exchange specification, web servers for providing the interaction data, clients for data integration and visualization, and a registry that maintains a list of publicly available servers (Figure 29.3).

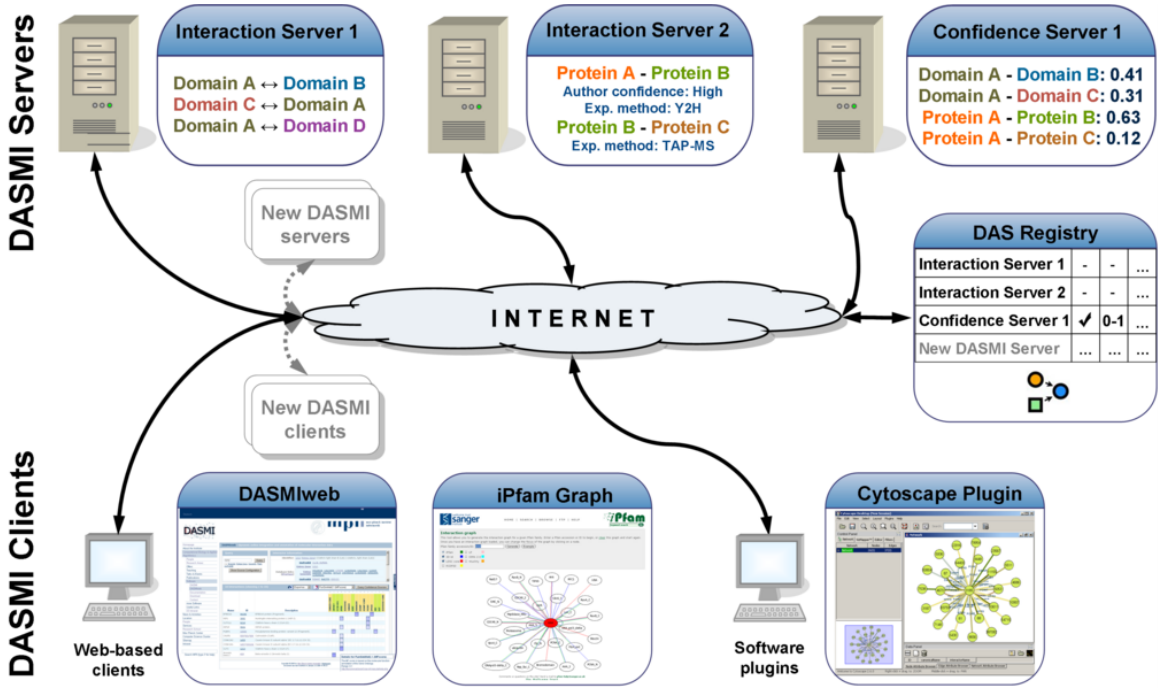


Figure 29.3: Schematic architecture of the Distributed Annotation System for Molecular Interactions. Interaction servers provide interactions (Interaction Server 1) and, optionally, additional information like experimental conditions (Interaction Server 2). Confidence servers provide quality scores for interactions (Confidence Server 1). DASMI clients query interaction and confidence servers and combine their results. The DAS registry maintains a list of available DAS servers.

We implemented different server types for the DASMI framework. *Interaction servers* provide interaction data and optionally additional information, such as the known or predicted interaction regions, the strength and type of the interaction, or the conditions under which the respective interaction occurs. *Confidence servers* supply quality scores for reported or hypothetical interactions. In this regard, DASMI provides the first confidence scoring system that is able to assess arbitrary sets of protein interactions. Since its realization is decentralized and thus not under the control of any particular authority, it is open for everyone to join. Scoring algorithms may be based on experimental or cellular conditions, co-expression, co-localization, functional co-annotation, evolutionary conservation, or topological network properties. Since no central authority can maintain all these different methods, the Proteomics

Standards Initiative (PSI) of the Human Proteome Organization (HUPO) concluded that our decentralized scoring architecture is preferable to central solutions [18].

Data exchange between servers and clients is managed by a DAS specification. To this end, we extended the original specification by the new *interaction* command and the associated DASINT XML response format, which are now part of the DAS 1.53E specification [15]. A great benefit of the distributed architecture is the instant extensibility by new servers and clients. To facilitate the process of setting up new servers, we extended the open-source DAS server libraries Dazzle (`http://www.biojava.org/wiki/Dazzle`) and ProServer (`http://www.sanger.ac.uk/Software/analysis/proserver/`). Further, to support the development of new DASMI clients, we upgraded the open-source DAS client libraries Bio-Das-Lite (`http://search.cpan.org/dist/Bio-Das-Lite/`) and Dasobert (`http://www.spice-3d.org/dasobert/`).

Using the aforementioned server libraries, we set up a variety of interaction servers. Our current setup comprises a dozen servers for protein-protein interaction datasets and twenty servers for domain-domain interactions of proteins. In addition, two confidence servers have been made available to score the reliability of protein interactions. One of the two servers assesses the functional similarity of the interacting proteins [22]. The other server determines whether a PPI can be traced to the underlying protein domain-domain interactions [20]. We expect from discussions with other scientists that new servers for molecular interactions and confidence scores will be made available at other institutions.

DASMI clients afford a convenient way for scientists to access, visualize, and analyze molecular interaction data provided by different servers. We developed the DASMI client DASMIweb (`http://www.dasmiweb.de`) as a user-friendly starting point for interactome studies. Another DASMI client for the free network visualization platform Cytoscape [25] is in preparation. DASMIweb presents interaction data in a clear tabular form for an intuitive, visual investigation of the results. In addition, interaction results can be exported in different formats, for example, for further analysis in external programs. Notably, the data pool of DASMIweb is fully configurable and allows the inclusion of additional, user-defined, servers.

We are currently cooperating with HUPO-PSI to develop the technology of the distributed architecture of DASMI further. This work consists of two subprojects, a system for distributed interaction data retrieval (PSICQUIC) and a system for distributed interaction confidence scoring (PSISCORE). The initial specifications are now available and DASMIweb and other DASMI clients will be fully compatible with them.

Integrative Search Engine and Data Mining

Data integration is a prerequisite for today's systems biology research [13] and is considered a critical step towards achieving a systems-level understanding of complex cellular processes [16]. Unfortunately, data integration continues to be a major challenge in bioinformatics due to the distributed organization of data sources, differing data quality, and the plethora of heterogeneous data representation schemes used. Although the proposed solutions to the problems are as diverse as the data sources, they tend to complement each other. The available solutions include cross-referencing datasets, for example, by SRS (`http://www.biowisdom.com/navigation/srs/srs`), data warehouses like BioMart (`http://www.biomart.org/`) and BioWarehouse (`http://biowarehouse.ai.sri.com/`) that store information in a central

repository, systems like DASMI (see above) that query various data sources and combine the results in a single view, and workflow-based frameworks like Taberna (`http://taverna.sourceforge.net/`) and BioMoby (`http://www.biomoby.org/`) that retrieve data from a number of data sources in a step-by-step procedure by following a workflow protocol defined by the user. In particular, the data warehousing model is the one that permits fast queries, retrieval, and transformation of integrated data once they are deposited and indexed in a local database.

Using this model, we have been building BioMyn (`http://www.biomyn.de`), a comprehensive data warehouse with a sophisticated web front-end for our systems biology projects (Figure 29.4) and cooperation partners with own datasets to be integrated. BioMyn is an ongoing project and still under development, but it already contains a comprehensive amount of integrated information on diverse biological annotations and relationships available for human genes and proteins from more than twenty external databases. The data warehouse stores almost three million annotations of diverse categories from Gene Ontology, sequence family classifications, protein domain architectures, metabolic and signaling pathways, protein interactions and protein complexes, disease associations, and UniProtKB keywords. Automatic updates from the source databases are carried out every month, and new data sources are added regularly. Updated annotations as well as annotations removed from the original sources are highlighted in our web site and are available through RSS feeds to facilitate the user's tracking of data changes. BioMyn also offers a number of online tools for analyzing groups of genes or gene products, for example, for computing the enrichment of certain Gene Ontology terms. The BioMyn search engine enables the biologist to find the original data source of an interesting annotation quickly, and the user can discover those genes or proteins that fulfill multiple criteria. For instance, the user may search for all proteins that are annotated as both being part of the plasma membrane and associated with a particular disease. In contrast, other data warehouses like BioMart or BioWarehouse do not offer complex search capabilities as BioMyn does, and web sites like the Bioinformatic Harvester (`http://harvester.fzk.de/harvester/`) act only as meta-search engine.

29.5.2 Modeling and Prediction

Investigators: Dorothea Emig, Sven-Eric Schelhorn, and Mario Albrecht

The Influence of Alternative Splicing on Protein and Domain Interaction Networks

In recent years, a large number of experimentally derived and computationally predicted protein-protein interactions (PPIs) and domain-domain interactions (DDIs) have become publicly available [19, 24, 1, 20]. It has also been discovered that more than 90 % of all human genes undergo alternative splicing [3, 31]. Therefore, alternative splicing may greatly influence protein diversity [28, 27]. Nevertheless, the existence of alternative splice variants is often neglected in current studies of protein interaction networks. However, protein isoforms may vary in their domain composition due to alternative splicing events and be involved with different DDIs underlying PPIs.

For this reason, we developed the plugin DomainGraph (`http://domaingraph.bioinf.mpi-inf.mpg.de/`) [6, 5] for the established open-source software Cytoscape [25], a platform for the visualization and analysis of molecular interaction networks. DomainGraph first

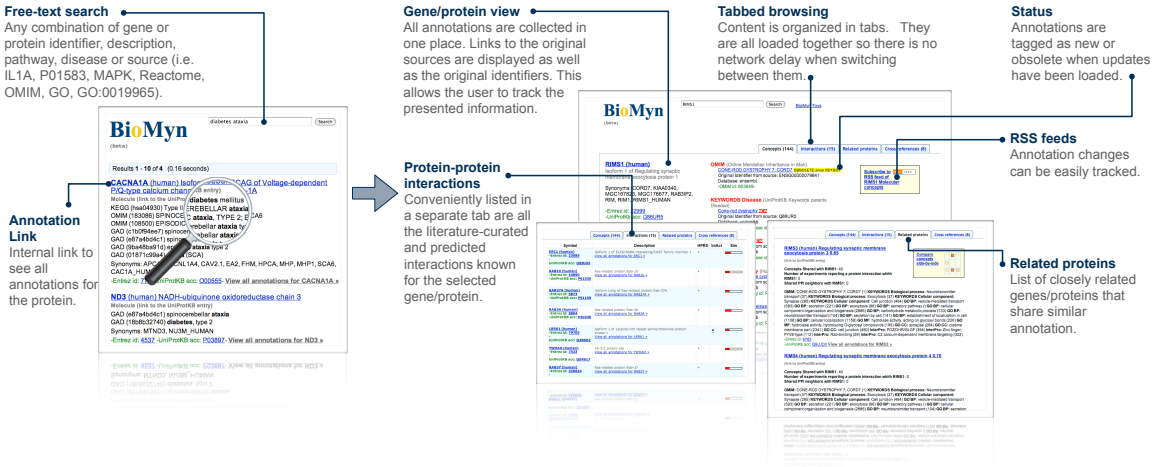


Figure 29.4: Web front-end of the BioMyn search engine with screen shots of the results page (left) and the annotation viewers (right).

decomposes the proteins of a user-imported gene or protein interaction network into their constituent protein domains (in case of a gene network, genes are first mapped to known protein isoforms) and then adds all known DDIs to the network (Figure 29.5). DomainGraph also provides further annotations about genes, proteins, and domains, for instance, Gene Ontology and OMIM disease annotations. In addition, DomainGraph can be used to visualize exon expression data and to highlight protein domains and their interactions affected by alternative splicing events (Figures 29.5 and 29.6). In the near future, DomainGraph will be extended to support the analysis of protein complexes as well.

Currently, twelve different species are supported including human, mouse, and yeast. For finding potential DDIs underlying imported PPIs, the user can select one out of over a dozen publicly available DDI datasets provided with DomainGraph. Our plugin also provides a mapping of Affymetrix Exon Array probesets to exons, Ensembl protein isoforms, and protein domains. This enables the user to import exon expression data and to overlay them with a gene or protein interaction network (Figure 29.5). In case of a single expression sample, the probesets are colored according to their presence or absence and their expression strength (Figure 29.6). For performing differential analysis with DomainGraph, we collaborate with the developers of the tool AltAnalyze (`http://www.genmapp.org/AltAnalyze`) at the University of California, San Francisco. For that purpose, the involved PhD student Dorothea Emig received a travel grant from the Boehringer Ingelheim Fonds for a research internship at this university from February to April 2009. The output of AltAnalyze from the comparison of paired exon expression samples can be used in DomainGraph to color up- and down-regulated probesets and corresponding protein domains.

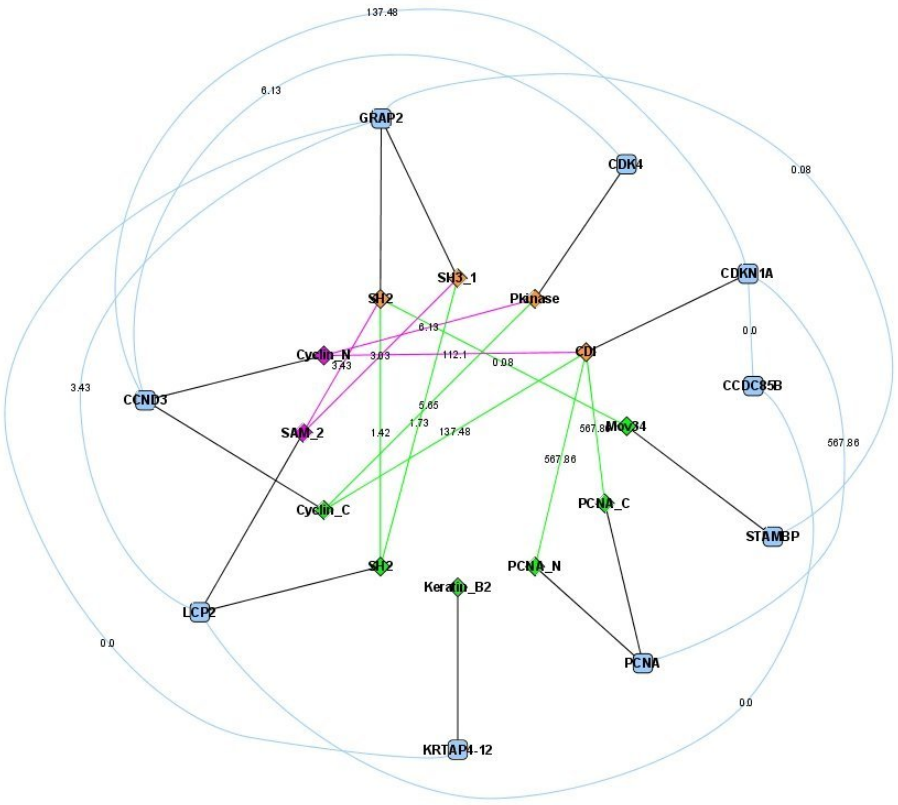


Figure 29.5: PPI and underlying DDI network with integrated exon expression data. Proteins are shown as rounded rectangles, while domains are represented by diamond-shaped nodes. Domains potentially spliced out according to the integrated expression data are colored pink, domains that are normally expressed are colored green, and domains that are indirectly affected (i.e. at least one of their interaction partners is spliced out) are colored orange. A special graph layout algorithm named RadialLayout is applied that combines elements from radial and layered algorithms [7]: protein and domain nodes are placed on two concentric circles, and an edge-crossing minimization procedure is used so that proteins are close to their constituent domains and PPI and DDI edges are separated visually; nodes of identical expression type are located close to each other.

Prediction of Interacting Protein Regions

Proteins are composed of modular protein regions that are associated with specific functions, and physical contacts between proteins are mediated by protein region interactions (PRIs) (Figure 29.7). The evaluation and prediction of PRIs are important for exploring protein-protein interactions (PPIs), analyzing protein function, structure and evolution, under-

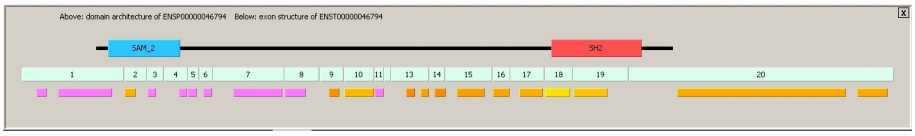


Figure 29.6: Graphical representation of a protein and the constituent domains (top row), the exon structure (middle) of the corresponding gene, and the annotated probesets (bottom) after integrating exon expression data. Probesets are colored according to presence (color gradient from yellow to red according to increasing expression strength) or absence (pink).

standing the assembly of protein complexes, and interpreting the effects of post-translational modifications in protein regions. Since only few interacting protein regions are known from atomic three-dimensional structures of interacting proteins, various computational methods have been developed to predict PRIs that explain protein interaction networks.

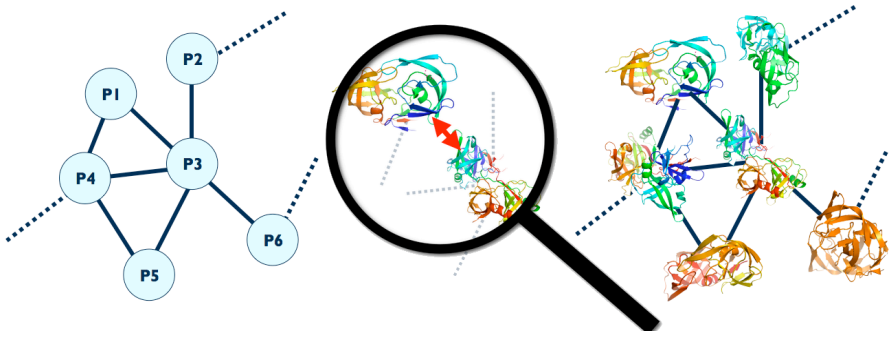


Figure 29.7: Experimentally observed protein-protein interactions (left) can be explained by interacting protein regions (middle), resulting in a network of interacting protein regions (right).

Three basic types of protein regions are known to be involved in PPIs: globular domains, peptides containing short linear motifs (SLiMs), and coiled coils. While interactions between the comparatively large and structurally well-conserved globular domains have been studied extensively and used by various prediction methods, PRIs between globular domains can only explain up to 20 % of all PPIs in major eukaryotic model organisms. Many PPIs are assumed to be mediated by either coiled coil regions or by SLiMs, especially in the case of transient PPIs. Unfortunately, the latter two region types are more difficult to investigate experimentally and thus were not taken in account by prediction methods for PRIs in the past. Since eukaryotic proteins tend to contain multiple protein regions, sophisticated cellular functions can also involve cooperative binding of multiple protein regions in interacting proteins.

In our recent publication [20], we introduced IPPRI (Integrative Prediction of PRIs) as an

integrative approach that applies a maximum likelihood method to predicting PRIs based on large datasets of pairwise PPIs as training data. Our approach does not only include all three region types (globular domains, SLiMs, and coiled coils), but can also consider confidence values for PPIs in the training data to down-weight less reliable PPIs. In addition, our method is able to deduce putative region cooperations in a computationally efficient manner. Furthermore, we employ one novel and two established scoring functions to improve the predictive power of our method. We also introduce a collection of validation methods to compare our method to gold standard PRIs obtained from crystal structures and small-scale biological experiments.

We find that all three region types can be utilized in a single prediction framework, resulting in many novel high-confidence predictions of the new region types SLiM and coiled coil as well as of region cooperations. Therefore, we could show the importance of SLiMs and coiled coils for explaining an observed protein interaction network, highlighting the need of further research focusing on these two region types. Apart from that, our novel scoring function that incorporates information about domain fusion events performs significantly better than previous scoring functions. We could also demonstrate that the inclusion of predicted high-confidence PPIs into the training dataset yields additionally predicted PRIs that are not discovered by other methods relying solely on experimentally detected PPIs as training data. Together with Elena Zotenko, we are currently extending our work to protein complexes to infer their inner structural topology regarding PRIs.

Resistance Analysis of Hepatitis C Virus

An estimated 150 – 200 million people worldwide live with an infection by the hepatitis C virus (HCV). This infection is also the leading cause for liver cancer and liver transplantations. As with HIV, no vaccine is available due to the very fast mutation rate of the virus. While the majority of patients can be cleared from the virus by HCV-unspecific drugs like interferon and ribavirin, the effect of these drugs varies strongly between patients and virus genotypes. Upcoming anti-HCV drugs (Figure 29.8) that act directly on viral proteins promise increased clearing rates in patients, but, similar to anti-HIV drugs, are prone to provoke escape mutations of the virus, which may result in drug-resistant virus quasi-species. As this phenomenon is well-known from HIV, it is our aim to transfer the bioinformatics methods developed for HIV resistance analysis to HCV. This work includes the analysis of mutation patterns, the interpretation of drug resistance, and the optimization of therapies using specific drugs individually or in combination.

Specifically, we plan to develop a comprehensive software package that will facilitate the fast prototyping of novel bioinformatics methods for virus analysis. This software package will be used for extending the existing geno2pheno system for HCV-relevant genotypic and phenotypic data. In particular, we plan not only to transfer existing models from HIV analysis to HCV, but also to develop new methods required for modeling phenomena especially relevant for HCV therapy like hepatocyte damage, cryoglobulinemia, cellular tropism, and virus-host protein interactions.

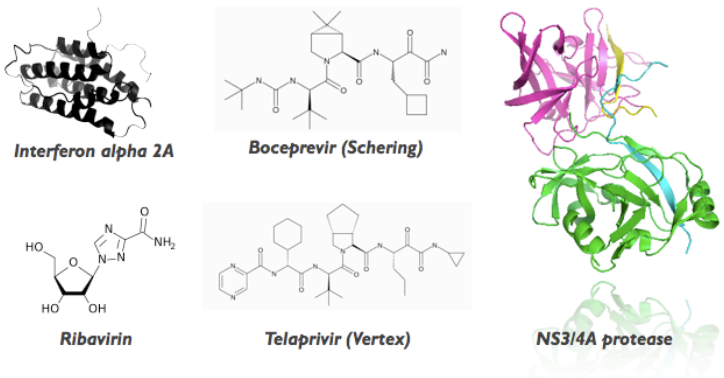


Figure 29.8: While interferon and ribavirin (left) are still the standard drugs administered against a hepatitis C virus infection, novel drugs (middle) specifically target viral proteins like the NS3/4A protease (right), thereby offering new treatment options.

29.5.3 Disease Focus

Investigators: Gabriele Mayr, Andreas Schlicker, and Mario Albrecht

We develop bioinformatics methods for identifying and characterizing disease genes and for advancing the systems biology understanding of disease processes and the effect of genetic variations on protein function and interactions as well as on drug therapies. To this end, we conduct application studies on the structure, function, and interaction of medically relevant proteins for deriving valuable biological hypotheses and suggesting further molecular wet-lab experiments by our cooperation partners. In addition, we devise novel disease candidate prioritization methods using the Gene Ontology and network data.

Hepatitis C Virus

In a recent study in close collaboration with Christoph Welsch from the Johann-Wolfgang-Goethe University Hospital of Frankfurt and Francisco Domingues [32], we have interpreted the effect of drug resistance developed by the hepatitis C virus (HCV) to the drug telaprevir, using a residue interaction network of the viral NS3-4A protease structure, see Section 29.7.4 for details. Further research work in the context of our DFG-funded Clinical Research Group on HCV (coordinated by Prof. Stefan Zeuzem at the Johann-Wolfgang-Goethe University Hospital of Frankfurt) is described in Section 29.5.2.

Auto-inflammatory Diseases

The intestinal immune system is tolerant to commensal bacteria as well as food antigens, while protecting the body against invading pathogens. This immune tolerance is disturbed in inflammatory bowel disease (IBD), which is a chronic auto-inflammatory disease with two

main subphenotypes, Crohn disease (CD) and ulcerative colitis (UC). IBD is a focus of our medical collaboration partners within the National German Research Network (NGFN) on Environmental Diseases coordinated by Prof. Dr. Stefan Schreiber at the University of Kiel.

Recently, their genome-wide association study implicated SNPs in NELL1, the gene for a cell-signaling protein controlling growth and differentiation, with both CD and UC [9]. The NELL1 protein is composed of an N-terminal thrombospondin (TSPN) domain and several subsequent von Willebrand factor (VWF) and epidermal growth factor (EGF) domains. Based on the crystal structure of the human thrombospondin-1 protein, we constructed a 3D model of the TSPN domain of NELL1 and localized the three SNPs associated with IBD. Our results indicate that local conformational changes induced by the SNPs in the TSPN domain may interfere with protein interactions abilities formed by NELL1. Another SNP resides in the VWF domain and might thus interfere with protein oligomerization.

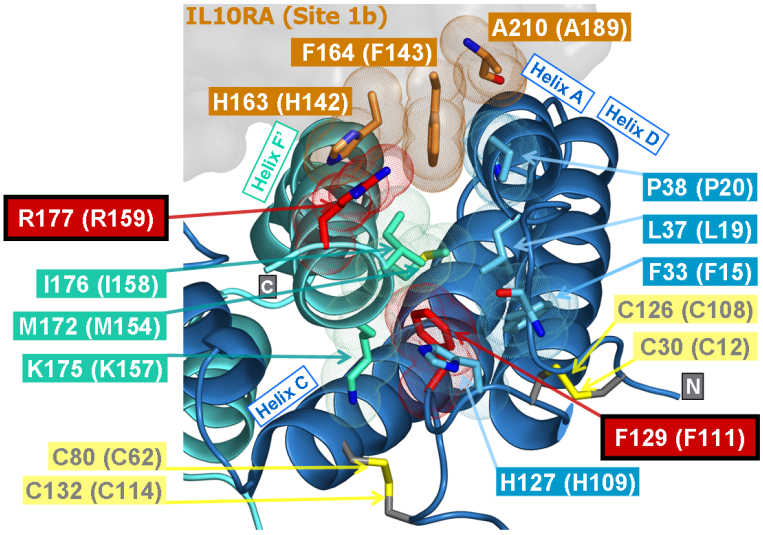


Figure 29.9: Close-up view of the structural environment around the two variant amino acids F129 and R177 (both colored red) inside the IL10 dimer. The two IL10 chains forming the intercalated IL10 dimer are colored blue and cyan, respectively. IL10 forms a receptor complex with IL10RA, whose surface is included in gray. Dotted transparent spheres illustrate van-der-Waals radii of selected amino acid atoms that form contacts. Amino acids in IL10 and IL10RA that are possibly affected by the variant F129 and R177 are shown as sticks. Atom coloring is as follows: oxygen in red, nitrogen in blue, sulfur in yellow. IL10 disulfide bonds are shown as gray and yellow sticks.

Furthermore, interleukin-10 (IL10), which has long been proposed to influence IBD pathophysiology, is another gene that our medical partners at the University of Kiel could confirm to be associated specifically with ulcerative colitis (UC) [8]. The IL10 locus encodes a protein that acts as an immunosuppressive cytokine and binds specific receptors to induce the tightly controlled JAK/STAT signaling pathway. The IL10-receptor interaction seems to be critical because the expression of viral homologs of the IL10 protein leads to the induction of an

alternative signaling pathway beneficial for virus invasion. Based on the crystal structures of human and viral IL10 proteins, we analyzed the effect of disease-associated SNPs in atomic detail (Figure 29.9). Various bioinformatics tools were applied to compare the related protein structures and to thoroughly study their binding interfaces. Interestingly, viral mimicry of IL10 concentrates on the structural interfaces between the cytokine and its receptors, while there is no relation between sequence conservation and receptor affinity. Both UC-associated SNPs found in the IL10 protein locate to a receptor binding site. Our results thus implicate that the IL10 immune response may be modified by SNPs, which affect receptor interaction and subsequently signaling pathways in the immune system, leading to inflammation.

Moreover, SNPs in the COL29A1 gene are associated with atopic dermatitis (AD), as discovered by our cooperating group directed by Prof. Dr. Young-Ae Lee at the Charité University Hospital in Berlin [26]. AD is a chronic inflammatory skin disorder and a major manifestation of genetically inherited allergic disease. The COL29A1 gene encodes a novel member of the collagen family that has been identified by our partners due to the occurrence of a number of SNPs in the gene locus. Our structural bioinformatics analysis studied possible effects of the disease-associated SNPs on the function of the COL29A1 protein. The protein is composed of a central short collagen triple helix flanked by ten von Willebrand factor A (VWA) domains. While six variants map to VWA domains, four SNPs locate to the triple helix of the collagen domain, but do not affect the repeating consensus sequence critical for the stabilization of the helices.

Plant Disease Resistance Proteins

A major strategy of the innate immune system of plants is the limitation of pathogen invasion by localized cell death. This apoptotic reaction is triggered by the interaction between specific pathogen proteins and plant resistance (R) proteins. R proteins are members of a large family of signal transduction NTPases, whose mammalian relatives, named NLRs, share similar protein domain architectures and functional modules with plant R proteins. Since some NLRs have been discovered to be linked to IBD and other human diseases, new insights into the various actions of the plant immune system may even contribute to the understanding of human immune response.

R proteins contain a central nucleotide-binding domain that most likely acts as a conformational switch between the active and the inactive state of the molecule. The C-terminal LRR domain is known to harbor negative as well as positive regulatory functions. Our experimental partner lab led by Dr. Frank Takken at the University of Amsterdam aims at elucidating the molecular mechanisms of R protein activation and downstream signaling. In particular, they focus on inter-domain interactions and the functional role of conserved sequence motifs. Those motifs are also conserved in the human protein Apaf-1, whose recently solved crystal structure served us as a template for modeling the 3D structure of R proteins. Our model-based findings suggest an inactive ADP-bound conformation as it is the case with Apaf-1, and the 3D structural location explains the effect of loss- or gain-of-function mutations observed in experiments performed by our partners [30, 29].

Semantic Similarity Measures for Protein Function Comparison

Semantic similarity measures are used for determining the similarity of concepts in structured vocabularies [23], for example, concepts represented by terms in the Gene Ontology (GO). Functional similarity methods leverage these semantic measures for comparing the functions of two gene products, for instance, associated with the same disease. Despite the large number of applications of semantic and functional similarity [24, 19, 10], no comprehensive resource for similarity values existed. Therefore, we developed the database FunSimMat (Functional Similarity Matrix, `http://www.funsimmat.de`) [22]. It provides a web front-end, a XML-RPC interface, and a DASMI web server allowing for easy manual and automatic access.

We implemented several semantic and functional similarity measures that are precomputed for proteins in UniProtKB, and protein families in Pfam and SMART. FunSimMat contains more than six million proteins and protein families, which results in roughly 36 trillion possible pairwise comparisons. In order to reduce this complexity, we introduced the concept of annotation classes. Each class consists of a lexically ordered set of GO terms from one GO ontology (biological process, molecular function, or cellular component). Functional similarity values are computed for all pairs of annotation classes, and each protein and protein family is annotated with the corresponding annotation classes. Furthermore, we also introduced a new functional similarity measures, the *rfunSimAll* score, that takes all three ontologies into account. FunSimMat has been available online since November 2007 and has already been queried over 1.4 million times. [1] To further improve the web service, we plan to implement additional similarity measures and functionalities like the MedSim prioritization methods described in the following.

Improving Disease Candidate Prediction using the Gene Ontology

More than 1,800 hereditary disorders are known to be caused by mutations in a single human gene [17]. However, susceptibility to many complex diseases like cancer, diabetes, and cardiovascular diseases is influenced by alterations in up to several hundred genes [12]. Therefore, computational techniques are being developed for prioritizing putative disease genes identified by genomic experiments and genome-wide association studies for further experimental verification [14].

We devised the new MedSim method for the comparison of disease phenotypes based on the functional Gene Ontology (GO) annotation of gene products that are known to be associated with the disease of interest. This new method allows for comparing phenotypes with each other and for comparing single proteins to phenotypes. Our results indicate that our method greatly supports the discovery of disease genes and performs well in prioritizing potential candidate genes.

References

- [1] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, 2008.

[1] This high number of hits results partly from whole-genome screens that are run by several users. We have had over 190 different users of FunSimMat.

- [2] H. Blankenburg, R. D. Finn, A. Prlić, A. M. Jenkinson, F. Ramírez, D. Emig, S.-E. Schelhorn, J. Büch, T. Lengauer, and M. Albrecht. DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10):1321–1328, 2009.

[3] B. J. Blencowe. Alternative splicing: new insights from global analyses. *Cell*, 126(1):37–47, 2006.

[4] R. D. Dowell, R. M. Jokerst, A. Day, S. R. Eddy, and L. Stein. The Distributed Annotation System. *BMC Bioinformatics*, 2:7, 2001.

- [5] D. Emig, M. S. Cline, K. Klein, A. Kunert, P. Mutzel, T. Lengauer, and M. Albrecht. Integrative visual analysis of the effects of alternative splicing on protein domain interaction networks. *Journal of Integrative Bioinformatics*, 5(2):101–115, 2008.

- [6] D. Emig, M. S. Cline, T. Lengauer, and M. Albrecht. Integrating expression data with domain interaction networks. *Bioinformatics*, 24(21):2546–2548, 2008.

- [7] D. Emig, K. Klein, A. Kunert, P. Mutzel, and M. Albrecht. Visualizing domain interaction networks and the impact of alternative splicing events. In *Proceedings of the 5th International Conference on BioMedical Visualization (MediVis): Information Visualization in Medical and Biomedical Informatics*, London, United Kingdom, 2008, pp. 36–43. IEEE Computer Society.

- [8] A. Franke, T. Balschun, T. H. Karlsen, J. Sventoraityte, S. Nikolaus, G. Mayr, F. S. Domingues, M. Albrecht, M. Nothnagel, D. Ellinghaus, C. Sina, C. M. Onnie, R. K. Weersma, P. C. F. Stokkers, C. Wijmenga, M. Gazouli, D. Strachan, W. L. McArdle, S. Vermeire, P. Rutgeers, P. Rosenstiel, M. Krawczak, M. H. Vatn, the IBSEN study group, C. G. Mathew, and S. Schreiber. Sequence variants in IL10, ARPC2, and multiple other loci contribute to ulcerative colitis. *Nature Genetics*, 40(11):1319–1323, 2008.

- [9] A. Franke, J. Hampe, P. Rosenstiel, C. Becker, F. Wagner, R. Häsler, R. D. Little, K. Huse, A. Ruether, T. Balschun, M. Wittig, A. ElSharawy, G. Mayr, M. Albrecht, N. J. Prescott, C. M. Onnie, H. Fournier, T. Keith, U. Radelof, M. Platzer, C. G. Mathew, M. Stoll, M. Krawczak, P. Nürnberg, and S. Schreiber. Systematic association mapping identifies NELL1 as a novel IBD disease gene. *PLoS ONE*, 2(8):e691.1–13, 2007.

[10] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78(6):1011–1025, 2006.

- [11] D. Frishman, M. Albrecht, H. Blankenburg, P. Bork, E. D. Harrington, H. Hermjakob, L. J. Jensen, D. A. Juan, T. Lengauer, P. Pagel, V. Schachter, and A. Valencia. Protein-protein interactions: analysis and prediction. In D. Frishman and A. Valencia, eds., *Modern Genome Annotation: The Biosapiens Network*, pp. 353–410. Springer, Wien, Austria, 2009.

[12] G. Gibson. Decanalization and the origin of complex disease. *Nat Rev Genet*, 10(2):134–140, 2009.

[13] C. Goble and R. Stevens. State of the nation in data integration for bioinformatics. *J Biomed Inform*, 41(5):687–693, 2008.

[14] T. Ideker and R. Sharan. Protein networks in disease. *Genome Res*, 18(4):644–652, 2008.

- [15] A. M. Jenkinson, M. Albrecht, E. Birney, H. Blankenburg, T. Down, R. D. Finn, H. Hermjakob, T. J. P. Hubbard, R. C. Jimenez, P. Jones, A. Kähäri, E. Kulesha, J. R. Macías, G. A. Reeves, and A. Prlić. Integrating biological data - the Distributed Annotation System. *BMC Bioinformatics*, 9(Suppl. 8):S3.1–7, 2008.

[16] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[17] T. P. O'Connor and R. G. Crystal. Genetic medicines: treatment strategies for hereditary disorders. *Nat Rev Genet*, 7(4):261–276, 2006.

[18] S. Orchard, J.-P. Albar, E. W. Deutsch, P.-A. Binz, A. R. Jones, D. Creasy, and H. Hermjakob. Annual Spring Meeting of the Proteomics Standards Initiative 23-25 April 2008, Toledo, Spain. *Proteomics*, 8(20):4168–4172, 2008.

• [19] F. Ramírez, A. Schlicker, Y. Assenov, T. Lengauer, and M. Albrecht. Computational analysis of human protein interaction networks. *Proteomics*, 7(15):2541–2552, 2007.

• [20] S.-E. Schelhorn, T. Lengauer, and M. Albrecht. An integrative approach for predicting interactions of protein regions. *Bioinformatics*, 24(16):i35–i41, 2008.

• [21] N. Scheller, P. Resa-Infante, S. de la Luna, R. P. Galao, M. Albrecht, L. Kaestner, P. Lipp, T. Lengauer, A. Meyerhans, and D. Juana. Identification of PatL1, a human homolog to yeast P body component Pat1. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1773(12):1786–1792, 2007.

• [22] A. Schlicker and M. Albrecht. FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Research*, 36(Database Issue):D434–D439, 2008.

[23] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7:1–16, 2006.

• [24] A. Schlicker, C. Huthmacher, F. Ramírez, T. Lengauer, and M. Albrecht. Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23(7):859–865, 2007.

[25] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, 2003.

• [26] C. Söderhäll, I. Marenholz, T. Kerscher, F. Rüschendorf, J. Esparza-Gordillo, M. Worm, C. Gruber, G. Mayr, M. Albrecht, K. Rohde, H. Schulz, U. Wahn, N. Hubner, and Y.-A. Lee. Variants in a novel epidermal collagen gene (COL29A1) are associated with atopic dermatitis. *PLoS Biology*, 5(9):e242.1952–1961, 2007.

[27] M. L. Tress, B. Bodenmiller, R. Aebersold, and A. Valencia. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol*, 9(11):R162, 2008.

• [28] M. L. Tress, P. L. Martelli, A. Frankish, G. A. Reeves, J. J. Wesselink, C. Yeats, P. Í. Ólason, M. Albrecht, H. Hegyi, A. Giorgetti, D. Raimondo, J. Lagarde, R. A. Laskowski, G. López, M. I. Sadowski, J. D. Watson, P. Fariselli, I. Rossi, A. Nagy, W. Kai, Z. Størling, M. Orsini, Y. Assenov, H. Blankenburg, C. Huthmacher, F. Ramírez, A. Schlicker, F. Denoued, P. Jones, S. Kerrien, S. Orchard, S. E. Antonarakis, A. Reymond, E. Birney, S. Brunak, R. Casadio, R. Guigo, J. Harrow, H. Hermjakob, D. T. Jones, T. Lengauer, C. A. Orengo, L. Patthy, J. M. Thornton, A. Tramontano, and A. Valencia. The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences*, 104(13):5495–5500, 2007.

• [29] G. Van Ooijen, G. Mayr, M. Albrecht, B. J. C. Cornelissen, and F. L. W. Takken. Transcomplementation, but not physical association of the CC-NB-ARC and LRR domains of tomato R protein Mi-1.2 is altered by mutations in the ARC2 subdomain. *Molecular Plant*, 1(3):401–410, 2008.

• [30] G. Van Ooijen, G. Mayr, M. M. A. Kasiem, M. Albrecht, B. J. C. Cornelissen, and F. L. W. Takken. Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. *Journal of Experimental Botany*, 59(6):1383–1397, 2008.

[31] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

• [32] C. Welsch, F. S. Domingues, S. Susser, I. Antes, C. Hartmann, G. Mayr, A. Schlicker, C. Sarrazin, M. Albrecht, S. Zeuzem, and T. Lengauer. Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of HCV. *Genome Biology*, 9(1):R16, 2008.

29.6 Computational Epigenetics

Coordinator: Christoph Bock

Epigenetics is commonly defined as the *study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence* [20]. Its fundamental objective is to elucidate mechanisms of gene regulation beyond transcription factor binding and cis-regulatory elements encoded in the DNA sequence. Epigenetic research touches upon two central problems of biology: How do cells specialize when a complex multicellular organism develops from a single fertilized egg [18]? And which molecular mechanisms contribute to phenotypic inheritance [19]?

Epigenetic mechanisms such as DNA methylation and histone modifications influence gene expression by modulating the packaging of the DNA in the nucleus. Patterns of epigenetic information are faithfully propagated over multiple cell divisions, making epigenetic regulation a key mechanism for cellular differentiation and cell fate decisions. In addition, incomplete erasure of epigenetic information can lead to complex patterns of non-Mendelian inheritance. Stochastic and environment-induced epigenetic defects are known to play a major role in cancer [2, 9, 11] and are likely to contribute to mental disorders [10, 17] as well as inflammatory diseases [1, 16, 21].

Recent developments, such as the ChIP-on-chip and ChIP-seq technologies, have started to convert epigenetic research into a high-throughput endeavor, to which bioinformatics is expected to make significant contributions. In the emerging field of computational epigenetics (see [4] for a detailed review), three areas currently stand out as most relevant:

- **Epigenetic data analysis.** Various experimental techniques have been developed for genome-wide mapping of epigenetic information, based on progress in high-resolution tiling microarrays and next-generation sequencing. All of these methods generate large amounts of data and require efficient ways of data processing, quality control and analysis by bioinformatic methods.
- **Epigenome prediction.** A substantial amount of bioinformatic research has been devoted to the prediction of epigenetic information from characteristics of the genome sequence. Such predictions, often based on support vector machines [22, 24], can serve dual purpose. First, accurate epigenome predictions can substitute for experimental data, to some degree, which is particularly relevant for newly discovered epigenetic mechanisms and for species other than human and mouse. Second, prediction algorithms derive statistical models of epigenetic information from training data and can therefore act as a first step toward quantitative modeling of an epigenetic mechanism (a key step toward integrating epigenetic regulation into systems biology models).

– **Analysis of disease-related epigenetic changes.** The functional role of epigenetic defects for cancer and other diseases opens up new opportunities for improved diagnosis and therapy. This active area of research gives rise to two questions that are particularly amenable to bioinformatic analysis. First, given a list of genomic regions exhibiting epigenetic differences between disease cells and healthy controls (or between different disease subtypes), can we detect common patterns or find evidence of a functional relationship of these regions to the disease? Second, can we use bioinformatic methods in order to improve diagnosis and therapy by detecting and classifying important disease subtypes?

Current research in the Computational Epigenetics Group aims to contribute to all three topics, by combining methods development, software implementation and collaborative application projects on roughly equal terms. In fact, these three approaches cross-fertilize each other to a significant degree: Newly developed methods are implemented into software packages, which can in turn lead to new applied science collaborations. Furthermore, working jointly with biomedical researchers on application projects is critical for identifying relevant problems for the next round of methods development, which closes the loop.

The main conceptual result of the report period was the discovery of an unexpectedly high correlation between genome and epigenome, using machine learning methods [6, 7, 8]. This result came as a surprise for many molecular biologists and has – in the meantime – been validated by multiple independent studies.

On the applied side, we have contributed to multiple biomedical studies [12, 13, 14, 15], and we have developed three software packages that are widely used by the epigenetic research community: (i) BiQ Analyzer (`http://biq-analyzer.bioinf.mpi-inf.mpg.de/`, [5]) has become a standard tool for processing DNA methylation data (dozens of analyses are performed every day by researchers from all over the world); (ii) EpiGRAPH (`http://epigraph.mpi-inf.mpg.de/`, [3]) enables biologists to analyze genome and epigenome datasets with powerful statistical and machine learning methods (hundreds of users before the official publication date); and (iii) MethMarker (`http://methmarker.mpi-inf.mpg.de/`, [23]) facilitates the design of DNA methylation assays and implements a systematic workflow for optimization and validation of DNA methylation biomarkers (part of our contribution to the EU-FP7 CancerDIP project).

The long-term goals of the Computational Epigenetics Group are to develop bioinformatic methods that facilitate the analysis and interpretation of epigenome datasets, and to use these methods in collaboration with biomedical and clinical researchers to improve epigenetic diagnostics and treatment. Currently, the main disease focus lies on cancer, for which the relevance of epigenetic alterations is well-established and large datasets are available. However, an accumulating body of evidence also suggests a relevant contribution of epigenetic regulation to diseases such as chronic inflammation [16, 21] and mental disorders [10, 17], providing exciting challenges for further research in computational epigenetics.

References

[1] I. M. Adcock, L. Tsaprouni, P. Bhavsar, and K. Ito. Epigenetic regulation of airway inflammation. *Current Opinion in Immunology*, 19(6):694–700, 2007.

[2] S. B. Baylin and J. E. Ohm. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nature Reviews Cancer*, 6(2):107–16, 2006.

• [3] C. Bock, K. Halachev, J. Büch, and T. Lengauer. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. *Genome Biology*, 10:R14, 2009.

• [4] C. Bock and T. Lengauer. Computational epigenetics. *Bioinformatics*, 24(1):1–10, 2008.

[5] C. Bock, T. Lengauer, S. Reither, T. Mikeska, M. Paulsen, and J. Walter. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics*, 21(21):4067–4068, 2005.

[6] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genetics*, 2(3):0243–0252, 2006.

• [7] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. CpG island mapping by epigenome prediction. *PLoS Computational Biology*, 3(6):1055–1070, 2007.

• [8] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Research*, 36(10):e55, 2008.

[9] M. Esteller. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, 8(4):286–98, 2007.

[10] A. P. Feinberg. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447(7143):433–40, 2007.

[11] A. P. Feinberg and B. Tycko. The history of cancer epigenetics. *Nature Reviews Cancer*, 4(2):143–53, 2004.

• [12] M. Kircher, C. Bock, and M. Paulsen. Structural conservation versus functional divergence of maternally expressed microRNAs in the Dlk1/Gtl2 imprinting region. *BMC Genomics*, 9:346, 2008.

• [13] F. Liu, E. Tostesen, J. K. Sundet, T.-K. Jenssen, C. Bock, G. I. Jerstad, W. G. Thilly, and E. Hovig. The human genomic melting map. *PLoS Computational Biology*, 3(5):e93, 2007.

• [14] T. Mikeska, C. Bock, O. El-Maarri, A. Hübner, D. Ehrentraut, J. Schramm, J. Felsberg, P. Kahl, R. Büttner, T. Pietsch, and A. Waha. Optimization of quantitative MGMT promoter methylation analysis using pyrosequencing and combined bisulfite restriction analysis. *Journal of Molecular Diagnostics*, 9(3):368–381, 2007.

• [15] D. Moser, S. Ekawardhani, R. Kumsta, H. Palmason, C. Bock, Z. Athanassiadou, K.-P. Lesch, and J. Meyer. Functional analysis of a potassium-chloride co-transporter 3 (SLC12A6) promoter polymorphism leading to an additional DNA methylation site. *Neuropsychopharmacology*, 34(2):458–467, 2009.

[16] E. L. Pearce and H. Shen. Making sense of inflammation, epigenetics, and memory cd8+ t-cell differentiation in the context of infection. *Immunological Reviews*, 211:197–202, 2006.

[17] A. Petronis. The origin of schizophrenia: genetic thesis, epigenetic antithesis, and resolving synthesis. *Biological Psychiatry*, 55(10):965–70, 2004.

[18] W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–32, 2007.

[19] E. J. Richards. Inherited epigenetic variation–revisiting soft inheritance. *Nature Reviews Genetics*, 7(5):395–401, 2006.

[20] V. E. A. Russo, R. A. Martienssen, and A. D. Riggs. *Epigenetic mechanisms of gene regulation.* Cold Spring Harbor monograph series ; 32. Cold Spring Harbor Laboratory Press, Plainview, N.Y., 1996.

[21] O. Sanchez-Pernaute, C. Ospelt, M. Neidhart, and S. Gay. Epigenetic clues to rheumatoid arthritis. *Journal of Autoimmunity*, 30(1-2):12–20, 2008.

[22] B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel methods in computational biology.* Computational molecular biology. MIT Press, Cambridge, Mass. ; London, 2004.

[23] P. Schüffler, T. Mikeska, T. Lengauer, and C. Bock. Optimized assay design for epigenetic biomarker development and high-throughput epigenotyping. submitted.

[24] V. N. Vapnik. *The nature of statistical learning theory.* Springer, New York, 1995.

29.6.1 Epigenome Analysis: Methods, Software, and Applications

Investigators: Konstantin Halachev and Christoph Bock

Our current work is centered around an epigenome analysis infrastructure provided by the EpiGRAPH software (`http://epigraph.mpi-inf.mpg.de/`, [1]). EpiGRAPH is a machine-learning driven web service for genome analysis and epigenome prediction. It was developed to enable biologists to perform complex bioinformatic analyses online – without the need to learn a programming language or to download large datasets. With its focus on statistical and machine learning methods, EpiGRAPH goes beyond existing tools, providing a comprehensive workflow that helps to uncover biologically meaningful associations among genome-scale datasets.

EpiGRAPH's value for epigenome research has been showcased by a number of studies that we conducted with early versions of the software [2, 3, 4, 6, 8]. The development of EpiGRAPH was initiated with the goal of predicting DNA methylation from properties of the genomic DNA sequence (e.g. DNA sequence and structure, genes and transcription factor binding sites). As it turned out, the support vector machines that are now at the heart of EpiGRAPH could predict DNA methylation with high accuracy, providing the first evidence that somatic CpG island methylation can be predicted computationally [2].

EpiGRAPH was later used in a follow-up study [3] to derive genome-wide predictions for a diverse set of epigenetic attributes, and we aggregated these predictions into an accurate annotation of bona fide CpG islands in the human genome. The main contribution of this project to EpiGRAPH was the proof of principle that genome-wide predictive analysis over multiple and diverse datasets (e.g. originating from different tissues and cell lines) is technically feasible and can be biologically productive.

Using a dataset from the European Human Epigenome Project [5], which comprises DNA methylation data for several dozen healthy individuals, we were able to quantify the degree of inter-individual variation of DNA methylation in a human population [4]. Furthermore, we could predict from the DNA sequence which genomic regions are prone to epigenetic variation and which regions are characterized by inter-individually stable DNA methylation patterns. This study provides results that help devise the most effective and cost-efficient strategy to experimentally map DNA methylation at high resolution in a large human cohort, which is a central goal of current epigenome projects.

Our most recent work uses EpiGRAPH to analyze epigenetic variation on developmental time-scales, performing DNA methylation prediction for different stages of cellular differentiation. A high-resolution, genome-scale analysis of tissue-specific DNA methylation has become possible only recently, utilizing bisulfite sequencing data for 5% of all CpG dinucleotides in the mouse genome that have been generated for more than a dozen cell types [7]. We analyzed and predicted DNA methylation profiles of several sets of functional genomic regions (including CpG islands, promoters and conserved elements) across all available cell types. Using EpiGRAPH, we showed that DNA methylation can be predicted with accuracies of above 75% across a wide range of tissues and developmental stages and for various genomic regions. Interestingly, prediction accuracies were consistently higher in pluripotent cells than in terminally differentiated cells, consistent with a recent hypothesis suggesting that DNA methylation of embryonic cells is coded in the DNA sequence and may form an epigenetic ground state that increasingly erodes as differentiation progresses [9].

References

- [1] C. Bock, K. Halachev, J. Büch, and T. Lengauer. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. *Genome Biology*, 10:R14, 2009.

[2] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genetics*, 2(3):0243–0252, 2006.

- [3] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. CpG island mapping by epigenome prediction. *PLoS Computational Biology*, 3(6):1055–1070, 2007.

- [4] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Research*, 36(10):e55, 2008.

[5] F. Eckhardt, J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38(12):1378–85, 2006.

- [6] F. Liu, E. Tostesen, J. K. Sundet, T.-K. Jenssen, C. Bock, G. I. Jerstad, W. G. Thilly, and E. Hovig. The human genomic melting map. *PLoS Computational Biology*, 3(5):e93, 2007.

[7] A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–70, 2008.

- [8] D. Moser, S. Ekawardhani, R. Kumsta, H. Palmason, C. Bock, Z. Athanassiadou, K.-P. Lesch, and J. Meyer. Functional analysis of a potassium-chloride co-transporter 3 (SLC12A6) promoter polymorphism leading to an additional DNA methylation site. *Neuropsychopharmacology*, 34(2):458–467, 2009.

[9] J. Silva and A. Smith. Capturing pluripotency. *Cell*, 132(4):532–6, 2008.

29.6.2 Cancer Epigenetics

Investigators: Yassen Assenov and Christoph Bock

Tumor development is driven by complex patterns of genetic and epigenetic abnormalities. For example, increased levels of DNA methylation at specific regulatory regions (e.g. promoters) can induce silencing of tumor suppressor genes, whereas decreased methylation levels in large non-genic regions contributes to genome instability. New methods such as MeDIP and ChIP-seq enable genome-wide comparison of epigenetic patterns between cancer and control samples. These technologies give rise to a bioinformatic challenge – how to analyze differences between cancer and control samples that are not restricted to genes, but can occur anywhere in the genome? We have developed two computational methods that address this problem.

The idea behind the first approach is to translate sets of genomic regions into sets of genes that are likely to be regulated by these regions. The associated genes are then analyzed with well-established methods such as Gene Ontology enrichment analysis. This method for estimating sets of affected genes is motivated by the increasing number of published genome-wide mapping studies, which produce abundant data on a variety of gene regulatory mechanisms, such as epigenetic modifications and protein binding. In order to identify targeted genes, our method systematically assesses the average locality of their regulatory effect. It relies on the assumption that most regulatory mechanisms target a subset of genes with characteristic biological functions. To that end, we have developed the Regions2Genes software, which assigns sets of genes to a given set of regulatory elements. A distinguishing feature of our method is that it takes chromosome structure into account, modeling the effect of insulators and epigenetic boundary elements. Using our method, we derived a biologically plausible estimate of the locality of Polycomb Repressive Complex 2 binding in human. We are currently extending this work to several transcription factors. As a complementary approach, we use the EpiGRAPH web service to identify common characteristics of genomic regions that are epigenetically altered in cancer cells.

These methods can assist in identifying new genomic regions with relevance for cancer development. However, it is equally important – and bioinformatically challenging – to develop a genomic region with an established role in cancer into an accurate and robust method for cancer diagnostics. We have addressed this topic in a pilot study focusing on the most deadly type of brain tumor, named glioblastoma multiforme. It is well-established that DNA methylation in the promoter region of the MGMT gene predicts chemotherapy resistance, based on a straightforward mechanism: The MGMT gene encodes a DNA repair protein, which corrects the specific type of DNA damage that is artificially induced by alkylating chemotherapy. Hence, when the MGMT gene is epigenetically silenced in a tumor, a large number cancer cells die at a cytotoxic dose that is still tolerated by normal cells (which can use their functional copy of the MGMT gene to repair chemotherapy-induced damage). On the other hand, when the MGMT gene is active in the tumor, cancer cells can effectively protect themselves against chemotherapeutic treatment. In this latter case, alkylating chemotherapy is rendered ineffective for patient survival and thus becomes an unnecessary strain to the patient. Together with our colleagues at the University of Bonn Medical Center, we have developed a method that effectively assesses the DNA methylation status of the MGMT gene promoter and statistically calculates the likelihood with which a given patient will respond to alkylating chemotherapy, based on a surgically obtained biopsy of the brain tumor [1]. With

this tool in hand, the treating oncologist can make an informed decision about whether or not chemotherapy is recommended for a given patient.

In the context of an ongoing EU-funded research project (`http://www.cancerdip.eu/`), we will for the first time combine our approaches for identifying new genomic regions with a role in cancer and for developing a subset of these regions into powerful cancer diagnostics tools. To be able to scale up our biomarker optimization method from a single gene to a sizable number of genomic regions, we have formalized and implemented the optimization procedure into a newly developed software tool (`http://methmarker.mpi-inf.mpg.de/`).

References

- [1] T. Mikeska, C. Bock, O. El-Maarri, A. Hübner, D. Ehrentraut, J. Schramm, J. Felsberg, P. Kahl, R. Büttner, T. Pietsch, and A. Waha. Optimization of quantitative MGMT promoter methylation analysis using pyrosequencing and combined bisulfite restriction analysis. *Journal of Molecular Diagnostics*, 9(3):368–381, 2007.

29.6.3 Evolutionary Epigenetics and Comparative Epigenomics

Investigators: Lars Feuerbach and Christoph Bock

Model organisms such as mice and rhesus macaques play a key role in the study of human diseases, but little is known about how closely their epigenetic regulation matches that of humans. As epigenetics-based methods for diagnosis and therapy become increasingly important, the interest into comparative epigenomics is growing.

The evolution of epigenetic modifications and the study and comparison of today's epigenomes are two sides of the same coin. A deeper understanding of their evolutionary history can explain the similarities and differences of current epigenomes, while detailed maps of epigenetic modifications for a broad spectrum of species, cell types and developmental stages can yield insights into epigenome evolution.

Cross-species Analysis of DNA Methylation

Previously, we have reported that the methylation density of CpG islands from human lymphocytes is anti-correlated to the conservation of the local DNA sequence [2]. This result underlines the triangular nature of the relationship between epigenetics, genetics and evolutionary processes: First, DNA methylation at cytosines is known to introduce a mutational burden on the genomic DNA sequences (5-methyl-cytosine is rapidly degraded), which links epigenetics with genome evolution. Second, epigenetically active genomic regions co-localize with the promoter regions of genes, linking genetically and epigenetically functional regions. Finally, evolutionary selection acting on genes provides a link between genome evolution and genetically functional elements.

In an approach to quantify the conservation of epigenetic information on a genome-wide scale we generated an automated workflow based on multiple-whole-genome alignments and EpiGRAPH [2]. This computational pipeline enables the inclusion of homologous genomic regions for sequence-based prediction of DNA methylation and can be interpreted as the epigenetic component of the described interdependence triangle.

Evolution and Conservation of CpG Islands

CpG dinucleotides are the (almost) exclusive targets of DNA methylation in mammalians, providing the genetic basis of this epigenetic switch. They are strongly underrepresented in the genomes of most species that exhibit genome-wide DNA methylation. However, CpG dinucleotides are enriched in a class of genome regions called CpG islands, which are assumed to be key carriers of epigenetic information. We noticed that existing programs for CpG island annotation were not well-suited for automated cross-genome analysis of CpG islands. Especially, they lack robustness against small changes in the DNA sequence, which introduces unacceptably high levels of noise into comparative studies [1]. Hence, we have generalized the widely applied single-sliding-window approach into a multiple-sliding-window approach that considers not one but all possible window sizes. In consequence, this algorithm reduces the number of effective search parameters and stabilizes the annotation outcome. Furthermore, we implemented a runtime optimized version of this algorithm to make its use computationally feasible for whole-genome annotations.

References

- [1] C. Bock. *Computational Epigenetics - Bioinformatic methods for epigenome prediction, DNA methylation mapping and cancer epigenetics.* Phd thesis, Universität des Saarlandes, 2008.

- [2] L. Feuerbach. Towards comparative epigenomics: A pilot study on dna methylation of cpg islands on human chromosome 21. Masters thesis, Universität des Saarlandes, 2007.

29.7 Protein Structure and Function

Coordinators: Francisco Domingues and Ingolf Sommer

Structural protein models determined experimentally by X-ray crystallography and NMR spectroscopy play a central role in the investigation of the molecular basis of protein function. Therefore such models are extremely valuable for the investigation and understanding of life's processes.

We have developed several approaches for the analysis of protein structure/function relationships. Subsection 29.7.1 describes our work on structural descriptors. These descriptors are computational representations of functionally relevant parts of protein structures. Inference of molecular function based on sequence/structure relationships is described in subsection 29.7.2. Many proteins perform their function by interacting with other proteins. Subsection 29.7.3 describes a method for comparing interfaces between interacting proteins. Finally in subsection 29.7.4 we describe a few medically relevant applications.

We closely cooperate with other teams in the Department. In particular we collaborate with the molecular networks team on residue interaction networks and on medical applications (section 29.5). We also collaborate with the HIV team regarding structural analysis of viral proteins (section 29.4), and with the computational chemical biology team (section 29.8).

29.7.1 Structural Descriptors

Investigators: Oliver Sander, Francisco Domingues, and Ingolf Sommer

Any computational analysis of protein structures requires an appropriate representation of the relevant parts of protein structures. This holds in particular for the analysis of structure-function relationships.

Consequently, one focus has been the development of structural descriptors, i.e. computational representations of relevant parts of protein structures. Essential properties of these structural descriptors are the invariance under rotations and translations of the described structures, as well as the property that similar structural sites are mapped to similar representations. These properties ensure efficient comparison of pairs of descriptors and fast retrieval of similar descriptors from a large database for a given query. Furthermore, the vectorial representation is amenable to statistical learning and multivariate analysis techniques.

In contrast to the combinatorial matching methods geometric hashing and clique detection on correspondence graphs, structural descriptors shift the computational effort from the comparison stage into the preprocessing stage. However, it should be noted that the distinction between combinatorial and descriptor-based methods is rather fuzzy. Here, we use the term *structural descriptors* in the sense of preprocessed representations, which are very descriptive on their own.

We developed two types of structural descriptors: distributions of pairwise inter-atom distances[7], and moment invariants [10]. They were first developed and assessed independantly and later assessed and compared to other, existing descriptor methods [6].

Moment Invariant Descriptor

Moment invariants are a concept from computer vision for the representation of point sets or densities. Recently, we published the application of moment invariants as structural descriptors for protein-protein binding sites [10]. Initially designed as descriptors of 2D point sets [4], Mamistvalov extended the concept to three-dimensional data [5]. Moment invariants are functions of the geometrical moments of a point set that are invariant under certain transformations. Typically these are orthogonal coordinate transformations, but functions that are invariants under scaling, shear or blurring have been developed as well. The invariant functions are based on the theory of algebraic invariants. In this context, we investigated four invariants proposed in [5] and [3].

While in [10] we used the moment invariants of untyped atom sets, an extension to atom types is straightforward. Using the type definitions from Schmitt et al. [9], partitioning all present atoms based on their respective type yields five sets of functional atoms. Each of their shapes can be described by moment invariants in turn. However, the centralization of the point sets or densities needs to be performed based on a commonly chosen center, e.g. the center of mass of all atoms, irrespective of type. This avoids the centralization of the differently typed atoms to different locations, as their centers of mass might vary. Thus each type yields a four dimensional feature vector which can be combined into one twenty-dimensional vector.

Distance Distribution Descriptor

Distance distributions use internal pairwise distances to represent the spatial arrangement of a set of atoms. Computing all pairwise distances for a set of atoms yields a distance matrix where the rows as well as columns correspond to specific atoms in the atom set and entries in the matrix at position i, j correspond to the Euclidean distance between the atoms i and j in three-dimensional space. For comparison of two atom sets from two molecules, both sets are represented by a distance matrix. If the correspondences between atoms in both sets are known a priori, then both distance matrices can be compared directly. This approach has been used previously, for example in the STRuster method [1, 2].

However, in many cases obtaining correspondences of atoms from both sets of atoms is difficult and computationally expensive. Using distributions of distances [7] affords comparing two atom sets without knowing the correspondences. All pairwise distances from the two respective distance matrices are aggregated into a vector and are interpreted to be randomly sampled from some distribution function. The dependencies given in the structure of the matrix, especially that all distances in a single row correspond to distances between a single atom in one set to all atoms in the other set, are ignored explicitly. By ignoring these dependencies, matching becomes simpler as both matrices do not have to be aligned prior to comparison. On the other hand, ignoring these dependencies discards information and could therefore theoretically lead to mapping dissimilar sets of atoms onto similar distance distributions. To which extent this reduction of available information hampers practical usability depends on the concrete application and was evaluated for the general comparison of protein-protein binding sites.

Combinatorial Matching

We refer to the concept of using a structural descriptor for describing the spatial arrangement for a given set of residues of a predefined binding site as SDbsite. Additionally, we investigated approaches for scanning the surfaces of two proteins for similarities, if the binding patches are not known a priori or if the delineation of the boundaries of the binding patches has a high degree of uncertainty. The latter is referred to as SDpatches. See Figure 29.10 for a schematic overview of the two scenarios.

For matching two sets of patches against each other, a probabilistic model for significance assessment of a match and two computational strategies for score integration were developed: (1) simple non-exclusive greedy matching assigns each patch in one protein to its best hit in the other protein, regardless of relative spatial position and allowing multiple patches to match to a single patch in the other protein, and (2) clique-based matching that accounts for the relative spatial position of the matched patches.

Assessment

As descriptors for pre-defined binding sites (SDbsite) we extensively assessed four structural descriptors, namely distance distributions, moment invariants, radial shells, and spin images, on various data sets [6]. Among these sets are a) a set of 50 protein kinase binding sites and the binding sites of their respective partners and b) larger sets of domains interfacing domains from the same and from different SCOP families. For the selection of these binding

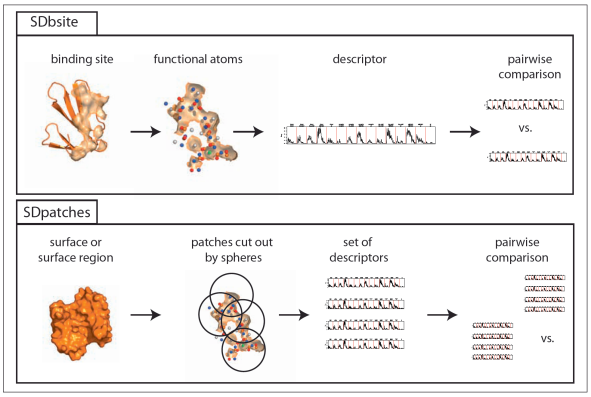


Figure 29.10: Schematic overview of SDbsite versus SDpatches

sites we used the SCOPPI database. Additionally, we performed tests on artificially distorted binding sites, where residues were subjected to random displacement or to deletion.

In the comparison of the four descriptors the distance distributions, radial shells, and spin images outperform moment invariants. However, in the evaluation on sites with simulated distortions, moment invariants exhibit advantages for distortions with a high proportion of deleted residues.

For describing surface patches (SDpatches) we restrict the analysis to using with distance distributions as the underlying base descriptor to represent the individual surface patches.

The performance was studied on the kinase data set and also in large scale retrieval experiments. For example on the kinase set, greedy patch-based matching of functional sites retrieves protein kinases with an accuracy of 71.05%. Including relative spatial information on the patches by clique-based matching further increases the accuracy to 76.32%.

These structural descriptors were further applied to the prediction of HIV coreceptor usage [8]. We anticipate further application in the context of function prediction as described in the following section.

References

[1] F. S. Domingues, J. Rahnenführer, and T. Lengauer. Automated clustering of ensembles of alternative models in protein structure databases. *Protein Engineering Design and Selection*, 17(6):537–543, 2004.

- [2] F. S. Domingues, J. Rahnenführer, and T. Lengauer. Conformational analysis of alternative protein structures. *Bioinformatics*, 23(23):3131–3138, 2007.

[3] J. Flusser, J. Boldis, and B. Zitova. Moment forms invariant to rotation and blur in arbitrary number of dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:234–245, 2003.

[4] M.-K. Hu. Visual pattern recognition by moment invatiants. *IRE Transactions on Information Theory*, 8:179–187, 1962.

[5] A. Mamistvalov. n-dimensional moment invariants and conceptual mathematical theory of recognition of n-dimensional data sets. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 20(8):819–831, 1998.

- [6] O. Sander. *Structural Descriptors for the Analysis of Protein Structure, Function, and Evolution.* Phd thesis, Universität des Saarlandes, 2008.
- [7] O. Sander, F. S. Domingues, H. Zhu, T. Lengauer, and I. Sommer. Structural descriptors of protein-protein binding sites. In A. Brazma, S. Miyano, and T. Akutsu, eds., *Proceedings of 6th Asia-Pacific Bioinformatics Conference*, Kyoto, Japan, 2008, pp. 79–88. Imperial College Press.
- [8] O. Sander, T. Sing, I. Sommer, A. J. Low, P. K. Cheung, P. R. Harrigan, T. Lengauer, and F. S. Domingues. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLoS Computational Biology*, 3(3):555–564, 2007.

[9] S. Schmitt, D. Kuhn, and G. Klebe. A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol*, 323(2):387–406, 2002.

- [10] I. Sommer, O. Müller, F. S. Domingues, O. Sander, J. Weickert, and T. Lengauer. Moment invariants as shape recognition technique for comparing protein binding sites. *Bioinformatics*, 23(23):3139–3146, 2007.

29.7.2 Function Prediction

Investigators: Ingolf Sommer, Oliver Sander, and Francisco Domingues

The GOdot system [7] assesses the level of conservation of function in regions of protein structure space. Protein structure space is defined in terms of several similarity measures for protein structures – currently CE [5], TM-align [8], and global and local profile alignments [3, 2] – which are used to precompute all-against-all similarities of 9500 reference protein domains. Each dimension of this space corresponds to the similarity to an individual reference domain. For four similarity measures and 9500 reference domains, we obtain a 38000 dimensional space. New proteins can be embedded into that space by relating them to the reference domains.

The molecular functions annotated to these protein domains are analyzed. Various regions in this space exhibit considerable difference in the local conservation of molecular function. The local conservation of function with respect to structure is modeled by means of logistic curves. Based on these models, we propose a method for predicting molecular function of a query protein with known structure but unknown function. The prediction method was rigorously assessed and compared with a previously published function predictor. Furthermore, we applied the method to 500 functionally unannotated PDB structures. The proposed approach provides a simple yet consistent statistical model for the complex relations between protein sequence, structure, and function. It has been tested extensively and successfully compared to state-of-the-art PHUNCTIONER [4] method. The GOdot method is available online (`http://godot.bioinf.mpiinf.mpg.de`).

Software Refactoring, Data Update, and Visualization

The GOdot system emerged from a Master's project [6]. In mid 2007, there was a preliminary web server, but the software behind was experimental. Parallel to writing the publication

[7], we undertook quite an effort to clean the code (a combination of R and Python, with a MySQL backend and an HTML and PHP frontend) and to keep the data maintainable.

Additionally, we worked on allowing to use current reference structures. While the publication is based on SCOP domains as representatives, our reference domains are now derived by mapping Pfam domain definitions onto known protein structures from the PDB.

We incorporate visualization of the structural and functional neighborhood of proteins, as follows: the precomputed all-against-all structural similarities of the reference proteins form a high-dimensional space, which can be reduced to two dimensions using multidimensional scaling techniques. The dimensionality reduction is performed locally, for user selected regions in the high-dimensional space. In particular the region around a query protein can be visualized. Additionally, information on the molecular function of the reference proteins can be mapped onto the two-dimensional representation. This affords a joint visualization of structural and functional information, which we think of as structure-function map around a query protein. An illustration is given in Figure 29.11A.

Study on Evolutionarily Defined Patches

We currently investigate the potential of adapting the function predictor GOdot towards incorporating patches of evolutionarily conserved protein residues computed by PatchFinder [1]. We assess this approach by comparing against standard GOdot. The method using the PatchFinder patches currently achieves a comparable performance. We analyze which of the PatchFinder patches really detect groups of functionally related proteins. With case studies on several Pfam families, we currently investigate the relation between patches and function; for an example see figure 29.12.

References

[1] G. Nimrod, F. Glaser, D. Steinberg, N. Ben-Tal, and T. Pupko. In silico identification of functional regions in proteins. *Bioinformatics*, 21:Suppl 1: i328–i337, 2005.

[2] N. Öhsen von, I. Sommer, R. Zimmer, and T. Lengauer. Arby: Automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics*, 20(14):2228–2235, 2004.

[3] N. von Öhsen, I. Sommer, and R. Zimmer. Profile-profile alignment: A powerful tool for protein structure prediction. In *Pacific Symposium on Biocomputing, PSB 2003*, Hawaii, USA, 2003, vol. 8, pp. 252–263. World-Scientific.

[4] F. Pazos and M. J. E. Sternberg. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A*, 101(41):14754–9, 2004.

[5] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11(9):739–747, 1998.

[6] N. Weinhold. Inference of protein function based on functionally conserved regions in sequence and structure space. Masters thesis, Universität des Saarlandes, 2006.

• [7] N. Weinhold, O. Sander, F. S. Domingues, T. Lengauer, and I. Sommer. Local function conservation in sequence and structure space. *PLoS Computational Biology*, 4:e1000105, 2008.

[8] Y. Zhang and J. Skolnick. TM-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Res*, 33(7):2302–2309, 2005.

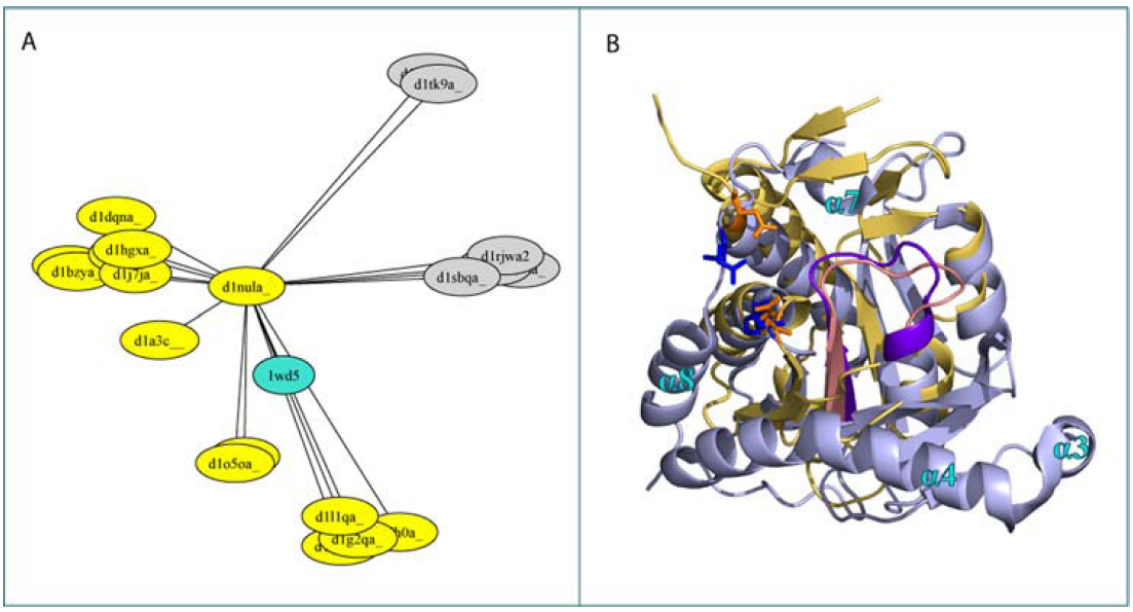


Figure 29.11: A) Hypothetical protein TT1426 (PDB 1wd5, colored turquoise) exhibits glycosyltransferase activity (GO:0016757). The structural neighborhood according to TM-align is visualized using multidimensional scaling. Proteins annotated with GO:0016757 are colored in yellow. They form a large group at the lower left section, where the query is also located. The glycosyltransferase group is subdivided into subgroups. In general these subgroups are associated with different substrates, in particular adenine phosphoribosyltransferase (d1l1qa, d1g2qa, clustering at the bottom), uracil phosphoribosyltransferase (d1o5oa, d1a3c, clustering squeezed between the other two subgroups), or xanthine/hypoxanthine/guanine phosphoribosyltransferases (d1nula, d1hgxa, d1dqna, d1j7ja, d1bzya, clustering at the left). Proteins not annotated with GO:0016757 are colored in grey. They are structurally related less closely to the query than the glycosyltransferases, and accordingly they group separately on the right and top. B) Structural superposition of query TT1426 (PDB 1wd5 in light blue) and the nearest neighbor, xanthine phosphoribosyltransferase (ASTRAL d1nula in gold).

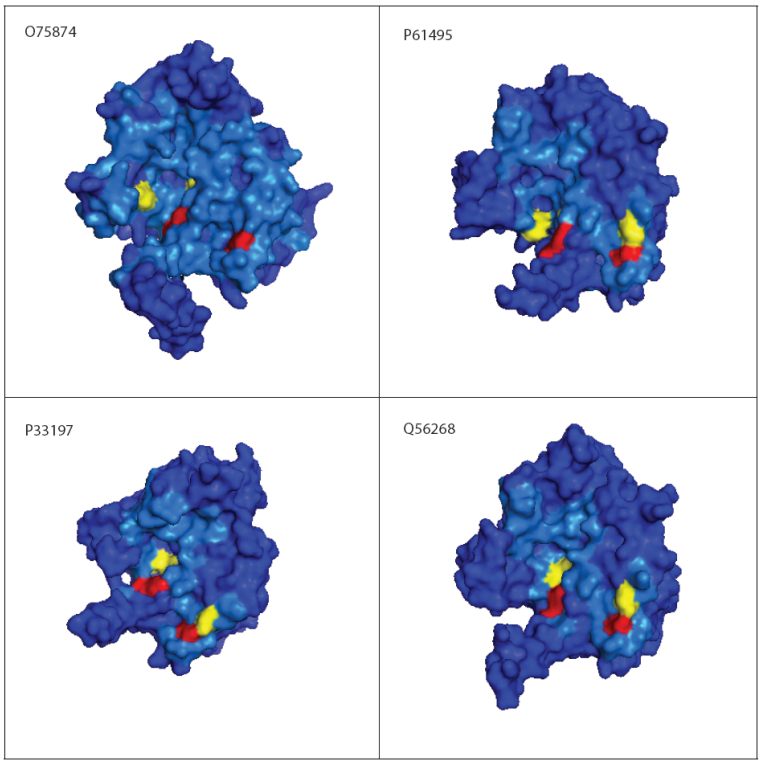


Figure 29.12: Structural patch visualization for the isocitrate/isopropylmalate dehydrogenases: The left-hand column shows two isocitrate dehydrogenases structures, the right-hand column shows two 3-isopropylmalate dehydrogenases structures. The patch residues are colored in light blue. The substrate binding sites as listed on the UniProtKB web site are highlighted in yellow. Catalytic sites from the CSA are shown in red.

29.7.3 Protein Interface Comparison

Investigators: Hongbo Zhu, Ingolf Sommer, and Francisco Domingues

Protein-protein interactions are involved in most cellular processes, and many proteins carry out their functions by forming complexes consisting of interacting polypeptide chains (subunits). The interfaces in such complexes are composed of complementary binding sites from the respective subunits. The analysis and comparison of protein-protein interfaces is essential for the understanding of the mechanisms of interaction between proteins. Such analysis is expected to have an impact on the prediction of interaction partners, as well as to assist in the design and engineering of protein interactions and interaction inhibitors.

We have previously characterized different types of protein-protein interaction types and developed an approach (NOXclass) for protein interface classification [1]. We have now developed Galinter, a method for aligning non-covalent interactions between different protein-protein interfaces [2]. The method aligns the vector representations of van der Waals interactions and hydrogen bonds based on their geometry, see Figure 29.13. The method is also applicable to comparing interfaces involving non-peptidic compounds. Two types of non-covalent interactions are considered: van der Waals interactions and hydrogen bonds. These interactions are represented as non-covalent interaction vectors (NCIVs) connecting the centers of two interacting atoms. The goal of the method is to find the largest set of NCIVs with similar geometric orientations. Two NCIVs (each from one interface) are matched in the alignment if they represent the same type of non-covalent interactions, have similar distances and relative orientations to the other matched NCIVs within the respective interfaces. A graph-based method is applied to matching the NCIVs. Galinter has been validated by comparison of interfaces involving homologous subunits. The alignments are consistent with the results obtained using complementary approaches that align binding sites and backbone atoms. A statistical model for assessment the significance of two aligned interfaces is currently under development.

Protein mimicry is relevant in the development of protein inhibitors. These inhibitors are designed such that their binding mode is similar to that of a wild-type protein-protein interaction. The development of new inhibitors is expected to benefit from detailed comparisons of the non-covalent interactions. We have applied Galinter to analyzing relevant cases of protein mimicry, see also subsection 29.7.4.

The Ser-His-Asp catalytic triad is present in many proteases and in different non-homologous protein families. In particular the trypsin-like serine proteases chymotrypsin and subtilisin are not homologous and lack sequence or structure similarity. Nevertheless they have been found to share as many as three types of inhibitors. We have analyzed the interactions formed between chymotrypsin and leech proteinase inhibitor eglin c, and subtilisin with chymotrypsin inhibitor 2. The two protease inhibitors have similar backbone structures and are homologous. It is noticeable that the NCIVs involving the catalytic serine and histidine residues are well conserved according to the Galinter alignment, see Figure 29.14.

References

[1] H. Zhu, F. S. Domingues, I. Sommer, and T. Lengauer. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7(27):1–15, 2006.

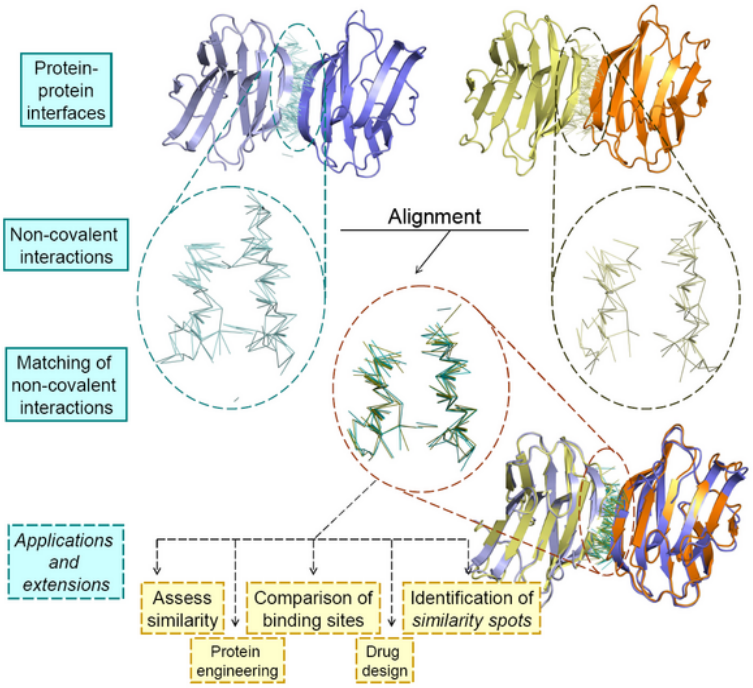


Figure 29.13: Interface alignment with Galinter. Non-covalent interactions are represented as vectors (NCIV) connecting the centers of interacting atoms. The vectors at the two interfaces are aligned using a graph based approach (finding maximum common subgraph). The resulting interface alignment is used to superimpose the two complexes.

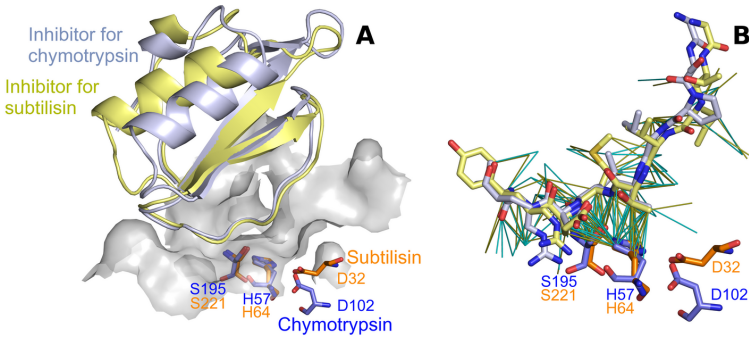


Figure 29.14: A) Superposed inhibitors and catalytic triads for chymotrypsin and subtilisin according to the Galinter alignment. The catalytic triads of chymotrypsin and subtilisin are shown as sticks. The molecular surface of the chymotrypsin active site is shown. B) superposed NCIVs for chymotrypsin/inhibitor interface and subtilisin/inhibitor interface according to the Galinter alignment. Only matched NCIVs are shown and the catalytic triad residues are labeled. The two catalytic triads are well superposed.

- [2] H. Zhu, I. Sommer, T. Lengauer, and F. S. Domingues. Alignment of non-covalent interactions at protein-protein interfaces. *PLoS ONE*, 3(4):e1926, 2008.

29.7.4 Medically Relevant Applications

Investigators: Hongbo Zhu, Christoph Welsch, and Francisco Domingues

Structural models determined experimentally provide unique and valuable insights in the investigation of the molecular basis of disease, as well as in the development of new drugs and therapeutic strategies. We have been performing structure-based analysis on several medically relevant problems involving viral and autoimmune diseases.

The HIV envelope glycoprotein gp120 binds to CD4 receptors located at the surface of target cells. We have investigated a CD4-mimetic antagonist (CD4M33-F23), that has been developed by engineering the gp120 binding site of CD4 onto a scorpion-toxin protein [2]. We have compared the natural CD4/gp120 interface with the antagonist/gp120 interface using Galinter [5]. In spite of the lack of similarity between the overall folds of CD4 and the antagonist, Galinter identifies extensive interface similarities indicating that the antagonist successfully mimics the CD4/gp120 interaction. In particular, about 80% of the non-covalent interactions at the antagonist/gp120 interface are aligned, including all interactions involving the hot spot residue Phe43 in CD4, see Figure 29.15.

Telaprevir is an inhibitor of the hepatitis C virus (HCV) protease NS3-4A, which is currently under investigation in phase 3 clinical trials. Previous clinical trials revealed residue mutations that confer varying degrees of drug resistance. In particular, mutations at two positions (V36 and T54) were associated with low to medium levels of drug resistance during viral breakthrough, together with only an intermediate reduction of viral replication fitness [3]. There is no obvious explanation for the molecular basis of drug resistance conferred by these mutations because they are located in the protein interior and far away from the ligand binding pocket. Together with the group of Prof. Dr. Stefan Zeuzem from the Johann Wolfgang Goethe University Hospital, Frankfurt and the team headed by Mario Albrecht (see also Section 29.5.2), we analyzed the network of non-covalent interactions in the neighborhood of the mutated positions in order to investigate the mechanisms of drug resistance [4]. In particular, we found that one of the mutants (T54A) is expected to affect hydrogen bonding between adjacent β-strands and as well as of a neighboring loop involved in shaping the inhibitor binding site (Figure 29.16). In addition, mutations at position V36 are expected to affect the conformation of a inhibitor-binding residue (F43), which may explain the increased drug resistance in the mutant forms.

Finally, in another collaboration with the research group of Mario Albrecht, we have investigated the molecular basis of ulcerative colitis, an inflammatory disease of the bowel, see Section 29.5.3) for details. In particular, we investigated the impact of IL10 sequence variants associated with the disease on the interaction with the corresponding interleukin receptor [1].

References

- [1] A. Franke, T. Balschun, T. H. Karlsen, J. Sventoraityte, S. Nikolaus, G. Mayr, F. S. Domingues, M. Albrecht, M. Nothnagel, D. Ellinghaus, C. Sina, C. M. Onnie, R. K. Weersma, P. C. F. Stokkers,

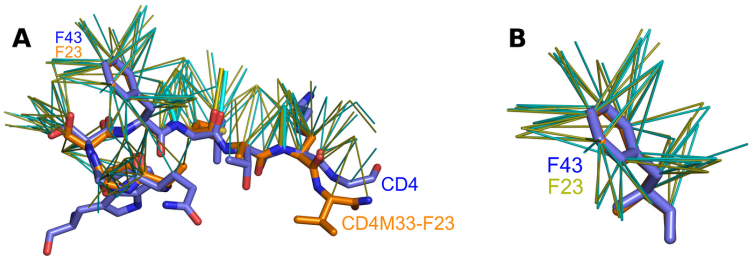


Figure 29.15: A) Superposed non-covalent interaction vectors (NCIVs) from the CD4/gp120 interface and CD4M33-F23/gp120 interface according to the Galinter alignment. CD4 is shown in dark blue and CD4M33-F23 is in orange. Only matched NCIVs are shown. CD4/gp120 NCIVs are shown in cyan, and CD4M33-F23/gp120 NCIVs are in yellow. Hydrogen bonds are shown as thick lines. B) An enlarged view of the matched NCIVs involving the hot spot phenylalanines.

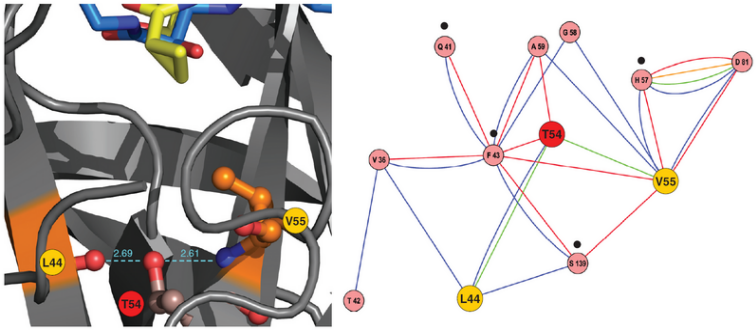


Figure 29.16: Investigating the molecular basis of viral drug resistance. Left: The HCV NS3-4A protease structure bound to two inhibitors CPX (yellow) and SCH 446211 (light blue). The H-bond interactions of T54 with L44 and V55 that bridge two β-strands are affected by the drug resistance mutation T54A. H-bonds are shown as lines and distances are given in angstroms. Right: network analysis of non-covalent residue interactions for T54 mutants. Residues that interact with the inhibitor are indicated by black dots. Nodes represent residues and colored edges represent different types of non-covalent interactions.

C. Wijmenga, M. Gazouli, D. Strachan, W. L. McArdle, S. Vermeire, P. Rutgeers, P. Rosenstiel, M. Krawczak, M. H. Vatn, the IBSEN study group, C. G. Mathew, and S. Schreiber. Sequence variants in IL10, ARPC2, and multiple other loci contribute to ulcerative colitis. *Nature Genetics*, 40(11):1319–1323, 2008.

[2] C. Huang, F. Stricher, L. Martin, J. M. Decker, S. Majeed, P. Barthe, W. A. Hendrickson, J. Robinson, C. Roumestand, J. Sodroski, R. Wyatt, G. M. Shaw, C. Vita, and P. D. Kwong. Scorpion-toxin mimics of cd4 in complex with human immunodeficiency virus gp120 crystal structures, molecular mimicry, and neutralization breadth. *Structure*, 13(5):755–768, 2005.

[3] C. Sarrazin, T. L. Kieffer, D. Bartels, B. Hanzelka, U. Müh, M. Welker, D. Wincheringer, Y. Zhou, H.-M. Chu, C. Lin, C. Weegink, H. Reesink, S. Zeuzem, and A. D. Kwong. Dynamic hepatitis c virus genotypic and phenotypic changes in patients treated with the protease inhibitor telaprevir. *Gastroenterology*, 132(5):1767–1777, 2007.

- [4] C. Welsch, F. S. Domingues, S. Susser, I. Antes, C. Hartmann, G. Mayr, A. Schlicker, C. Sarrazin, M. Albrecht, S. Zeuzem, and T. Lengauer. Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of HCV. *Genome Biology*, 9(1):R16, 2008.
- [5] H. Zhu, I. Sommer, T. Lengauer, and F. S. Domingues. Alignment of non-covalent interactions at protein-protein interfaces. *PLoS ONE*, 3(4):e1926, 2008.

29.8 Computational Chemical Biology

Coordinator: Iris Antes

The work in this section focuses on the development of new methods and application studies in the area of docking and drug design. The development of highly active drugs for efficient and safe treatment of diseases is the major goal of pharmaceutical research. Identifying new candidates for drug design is a very challenging process during which computational and experimental scientists work hand in hand.

Computational methods can aid the drug design process at several levels depending on the experimental data available. If the structure of the target protein or a closely related protein is known the most common approach used is molecular docking. Molecular docking methods are very efficient methods, which are specifically designed for rapid identification and characterization of protein-ligand interactions. The main application area of these methods is virtual drug screening.

However, virtual screening places tight CPU time constraints on the applied methodology due to the enormous number of potential drug molecules to be tested (around 100,000 to 1,000,000 molecules per screen). Due to these time constraints one very rough approximation is commonly used: The structural changes in the receptor upon ligand binding are neglected. Several well performing docking programs are meanwhile available. Our work focuses on the development of new and improved methods for settings for which the existing algorithms do not perform well.

One major project in the field of bio-molecular docking focuses on the efficient treatment of protein flexibility during docking and docking in situations for which an experimental structure of the receptor is not available and a theoretical structure obtained by homology modeling must be used. For this purpose the IRECS approach was developed. In a second project (DynaDock) we are working on a combined bioinformatics/biophysics based algorithm for docking peptide

and peptidic ligands into flexible binding sites (section 29.8.1). Traditional docking algorithms do not perform well for peptide docking, mainly because of the large number of degrees of freedom of the peptidic ligands and the mostly solvent exposed binding sites. This work is also related to and based on the experiences from a structural immuno-informatics project in which a combined sequence- and structure-based method (DynaPred) for the prediction of peptide binders to MHC-molecules was developed. The approach was extended to different MHC class I allele types (section 29.8.3). In addition to the methodological projects, several application studies were performed in the fields of docking (section 29.8.2) and structural immuno-informatics (section 29.8.3).

29.8.1 Docking into Flexible Proteins

Investigators: Iris Antes, Christoph Hartmann, Matthias Dietzen, and Elena Zotenko

FlexE-IRECS: Docking into Homology Models of Protein Structures

One prerequisite for performing *in-silico* prediction regarding the function and stability of proteins and their complexes is the availability of high quality protein structures. Although the number of experimentally solved protein structures is increasing rapidly, such information is still lacking for many important proteins. In these cases one has to rely on theoretically modeled structures built by homology modeling. However, the quality of homology modeled protein structures is seldom sufficient for scenarios in which the atomistic details of the models are important, like molecular docking. The one major reason for this is the limited accuracy especially of the predicted side-chain conformations in homology models. Using alternative side chain conformations during docking can improve the results considerably.

For this purpose the program IRECS [4] was developed. IRECS (Iterative REduction of Conformational Space) is a new tool for side-chain placement, which is especially tailored to the needs of molecular docking. In contrast to other side-chain placement tools, which predict the same number of conformations (mostly one) for all side chains of the protein, our tool is able to predict an ensemble of the most probable conformations for each side chain of a protein. The numbers of rotamers that are assigned to each side chain correspond to the flexibility of the respective side chain. This is a crucial feature for molecular docking, because the most successful strategy to deal with side-chain placement uncertainties in homology models during docking is the use of ensembles of alternative side-chain conformations in the binding pocket. IRECS provides an optimized set of side-chain conformations for this purpose.

We evaluated the program by generating both rigid and flexible protein models with our side-chain prediction tool IRECS and docked ligands to proteins using the newly developed scoring function ROTA and the docking programs FlexX [8] (for rigid side chains) and FlexE [3] (for flexible side chains) [5]. We validated our approach on the forty screening targets of the DUD database [6]. The validation shows that the ROTA potentials are especially well suited for estimating the binding affinity of ligands to proteins. The results also show that our procedure can compensate for the performance decrease in screening that occurs when using protein models with side chains modeled with a rotamer library instead of using X-ray structures. The average runtime per ligand of our method is 168 seconds on an Opteron V20z, which is fast enough to facilitate virtual screening of compound libraries for drug candidates.

DynaDock: Fully Flexible Protein-Peptide Docking

Next to the inclusion of receptor flexibility, the docking of very large, flexible ligands like peptides still poses a challenge for molecular docking programs [7]. We developed a new approach to docking peptide ligands into flexible receptors. For this purpose a two step procedure was developed: first, a broad sampling step is performed, which is used to scan the protein-peptide conformational space and to identify approximate ligand poses and second, in a fine sampling step, the ligand poses from step one are refined. For the fine sampling a new molecular dynamics based method was developed, called Optimized Potential Molecular Dynamics (OPMD), which is based on the self-consistent optimization of soft-core potentials [1]. Comparison with refinement results obtained by conventional molecular dynamics showed that refinement capabilities of the new method are significantly higher than for normal molecular dynamics approaches. This allows for a very efficient overall docking algorithm, because the necessary accuracy and thus the number of starting poses needed for successful refinement is much lower. In addition, due to the use of a molecular dynamics based approach we are able to allow for full flexibility of the whole system during the OPMD step. The algorithm was evaluated on a test set of 15 protein-peptide complexes with 2mer to 16mer peptides. Docking poses with an RMSD from the starting structure with less than 2.1Å were obtained in all cases. In addition, we fitted a simple interaction energy based scoring function. Applying this scoring function, in 11 out of the 15 cases the poses with the lowest score had a peptide RMSD smaller than 2.0Å. For the remaining three cases the RMSD values were between 2.0 and 3.54Å.

DynaCell

A new software package, DynaCell, was written specifically for the development and evaluation of DynaDock and especially the OPMD approach. Although the program includes an overlap guided random docking module, its focus is on the refinement of docked poses and thus poses from other docking tools like FlexX can also be imported for refinement. The software was merged with the IRECS program, such that the optimized side-chain conformations and homology models generated by IRECS can directly be used as an input to DynaCell. Since we aim at making the framework independent from third-party software, the functionality has been extended towards generating both side-chain conformations and protein topology files for OPLS-AA and GROMOS force fields, performing simple structure checks and adding hydrogens to proteins according to these force fields.

Fully Flexible Protein-Ligand Docking Incorporating Backbone Flexibility

The time-efficient prediction of conformational changes performed by a protein when binding a ligand is still a challenging yet crucial problem in computer-aided drug design. Indeed, the prediction of side-chain re-orientations for a given backbone conformation upon docking is feasible with sufficient precision and speed. Methods that can dock into an ensemble of (backbone) conformations, derived for example from PDB crystal structures or MD simulations, also exist. However, the problem of determining conformational changes of the backbone when exposed to a ligand in an accurate and time-efficient manner is much more complex and has thus hardly been tackled so far. Our aim is to investigate the mechanisms

that are involved in backbone motions upon ligand binding and gain insights which will provide the base for a docking method for Virtual Screening experiments that is able to effectively model these mechanisms.

References

[1] I. Antes. DynaDock: A molecular dynamics based method for protein-peptide docking including receptor flexibility. submitted.

- [2] I. Antes. *Computational Methods for the Investigation of Protein-Ligand Interactions*. Habilitation thesis, Universität des Saarlandes, 2009.

[3] C. Claussen, C. Buning, M. Rarey, and T. Lengauer. Flexe: Efficient molecular docking considering protein structure variations. *Journal of Molecular Biology*, 308(2):377–395, 2001.

- [4] C. Hartmann, I. Antes, and T. Lengauer. IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Science*, 16:1294–1307, 2007.

- [5] C. Hartmann, I. Antes, and T. Lengauer. Docking and scoring with alternative side-chain conformations. *Proteins: Structure, Function, and Bioinformatics*, 74(3):712–726, 2009.

[6] N. Huang, B. K. Shoichet, and J. J. Irwin. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801, 2006.

[7] Z. Liu, B. N. Dominy, and E. I. Shakhnovich. Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential. *Journal of the American Chemical Society*, 126(27):8515–28, 2004.

[8] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 261(3):470–89, 1996.

29.8.2 Application Studies

Investigators: Iris Antes and Christoph Hartmann

Several application studies were performed in close collaboration with experimental groups.

Inhibitor Design for Aldsterone Synthase

This project is a classical drug design project and has been performed in close collaboration with the group of Prof. Dr. R. Hartmann at the Pharmaceutical Chemistry Department of the Saarland University. The goal of the project was to find inhibitors for aldosterone synthase. The enzyme catalyzes the final steps of glucocorticoid (corticosterone) and mineralocorticoid (aldosterone) production. Overproduction of its product, aldosterone, is responsible for various cardiovascular deceases and thus its inhibition of wide-spread interest for the pharmaceutical industry.

At the start of the project several well inhibiting compounds for CYP11B2 had been identified in the group of Prof. Hartmann [6]. These compounds however, were not selective with respect to other steroid producing and metabolizing CYP P450 systems and did not have high binding affinities. Thus the goal of our project was to find highly potent, selective inhibitors for CYP11B2, which do not inhibit other CYP P450s significantly. Due to the

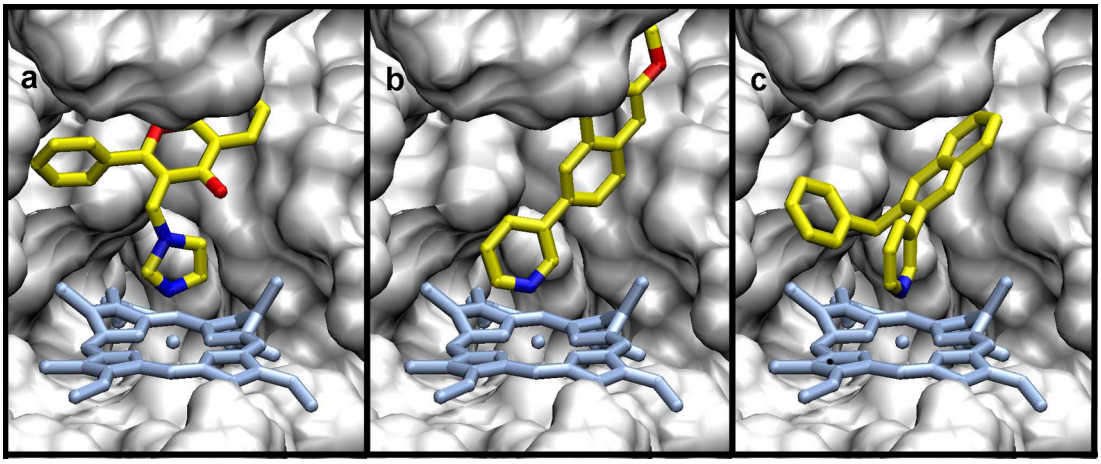


Figure 29.17: Structure of three CYP11B2-inhibitor complexes. Surface of the binding pocket (gray) surrounding the inhibitor and the heme cofactor (light blue). The inhibitors are presented in yellow; nitrogen atoms are colored in blue, and oxygen atoms are in red. In the inhibitor (c) was computationally designed on the basis of the binding modes of inhibitors (a) and (b).

difficulty of resolving membrane-binding proteins, there are no experimental structures available for the cytochromes CYP11B1 and CYP11B2. Using the recently resolved human cytochrome CYP2C9 structure (PDB code: 1R9O) as a template, we built 3D structural models for CYP11B2 and CYP11B1. Comparison of these models with two previously built structural models for CYP11B1 and CYP11B2 showed large differences in the models overall structures. These differences could be traced to the use of different template structures for the modeling process. This observation is in agreement with another study on CYP2D6, which also demonstrated the strong dependency of the targets structure on the templates used [2]. This demonstrates the limited accuracy of the modeled structures and thus the need of a further refinement of the models for successful docking.

We refined our model structures using the information we had about the known inhibitors for CYP11B2. The refinement was performed by alternating energy minimization/simulated annealing calculations and docking steps using the DynaCell program. We evaluated the refined protein structures by the docking of known inhibitors and non-inhibitors. We were able to increase the number of inhibitors which docked successfully into the binding pocket from 9.7% in the homology model to 90.3% in the refined model. At the same time the number of docked non-inhibitors stayed nearly the same at 20%. Using the refined protein models we were able to contribute significantly to the lead refinement process over the last years [3, 4, 6, 5, 2].

Rational Protein Design of CYP106A

Steroids are important pharmaceutically active compounds. In contrast to the liver drug-metabolizing cytochrome P450s, which metabolizes a variety of substrates, steroid hydroxy-

lases generally display a rather narrow substrate specificity. It is therefore a challenging goal to change their regio- and stereo-selectivity. CYP106A2 is one of only a few bacterial steroid hydroxylases and hydroxylates 3-oxo-delta4-steroids mainly in 15 beta-position.

In order to gain insights into the structure and function of this enzyme, whose crystal structure is unknown, a homology model has been created. The substrate progesterone was then docked into the active site to predict which residues might affect substrate binding. The model was substantiated by using a combination of theoretical and experimental investigations. First, numerous computational structure evaluation tools were applied to assess the plausibility of its protein geometry and its quality. Second, many key properties of common cytochrome P450s could be explained by the model. Third, two sets of mutants have been heterologously expressed, and the influence of the mutations on the catalytic activity towards deoxycorticosterone and progesterone has been studied experimentally: the first set comprises six mutations located in the structurally variable regions of this enzyme that are very difficult to predict by cytochrome P450 modelling (K27R, I86T, E90V, I71T, D185G and I215T). For these positions, no participation in the active-site formation was predicted, or could be experimentally demonstrated. The second set comprises five mutants in substrate recognition site 6 (S394I, A395L, T396R, G397P and Q398S). For these residues, participation in active-site formation and an influence on substrate binding was predicted by docking. These mutants are based on an alignment with human CYP11B1, and in fact most of these mutants altered the active-site structure and the hydroxylation activity of CYP106A2 dramatically.

Analysis of the Influence of Resistance Mutations in HCV Protease on Drug Binding

The inhibitor telaprevir (VX-950) of the hepatitis C virus (HCV) protease NS3-4A has been tested in a recent phase 1b clinical trial in patients infected with HCV genotype 1. This trial revealed residue mutations that confer varying degrees of drug resistance. In particular, two protease positions with the mutations V36A/G/L/M and T54A/S were associated with low to medium levels of drug resistance during viral breakthrough, together with only an intermediate reduction of viral replication fitness. These mutations are located in the protein interior and far away from the ligand binding pocket. Based on the available experimental structures of NS3-4A, we analyze the binding mode of different ligands. We analyzed the potential impact of V36 and T54 mutants on the side chain and backbone conformations and on the non-covalent residue interactions.

We propose possible explanations for their effects on the antiviral efficacy of drugs and viral fitness. For this purpose molecular dynamics simulations were performed of the T54A/S mutants and the IRECS approach was applied for rotamer analysis of the V36A/G/L/M side chain mutations. The results of these studies were combined with further analyses performed by the Protein Structure and Function Prediction group (see 29.7.4). Experimental data using an HCV V36G replicon assay corroborate our findings. T54 mutants are expected to interfere with the catalytic triad and with the ligand binding site of the protease. Thus, the T54 mutants are assumed to affect the viral replication efficacy to a larger degree than V36 mutants. Mutations at V36 and/or T54 result in impaired interaction of the protease residues with the VX-950 cyclopropyl group, which explains the development of viral breakthrough variants.

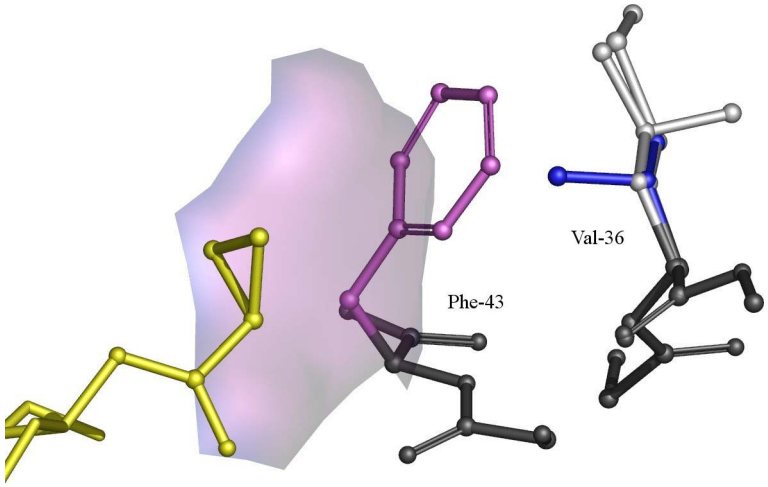


Figure 29.18: Rotameric states and conformational differences for V36 mutants A/G/L/M of the HCV protease NS3-4A computed with IRECS using the PDB entry 1RTL. The figure illustrates the relative position of mutant side chains (light grey) and the wild-type residue V36 (blue). Protein backbone changes are depicted in black. The contribution of F43 to the hydrophobic cavity conformation and the cyclopropyl binding pocket is illustrated by means of a transparent surface patch.

References

- [1] M. Lisurek, B. Simgen, I. Antes, and R. Bernhardt. Theoretical and experimental evaluation of a CYP106A2 low homology model and production of mutants with changed activity and selectivity of hydroxylation. *ChemBioChem*, 9(9):1439–1449, 2008.
- [2] S. Lucas, R. Heim, M. Negri, I. Antes, C. Ries, K. E. Schewe, A. Bisi, S. Gobbi, and R. W. Hartmann. Novel aldosterone synthase inhibitors with extended carbocyclic skeleton by a combined ligand-based and structure-based drug design approach. *Journal of Medical Chemistry*, 51(19):6138–6149, 2008.

[3] S. Ulmschneider, U. Mueller-Vieira, C. D. Klein, I. Antes, T. Lengauer, and R. W. Hartmann. Synthesis and Evaluation of Pyridylmethylene-tetrahydronaphthalenes and -indanes: Potent and Selective Inhibitors of Aldosterone synthase (CYP11B2). *Journal of Medicinal Chemistry*, 48(13):4489–4490, 2005.

[4] S. Ulmschneider, U. Mueller-Vieira, M. Mitrenga, R. W. Hartmann, S. Oberwinkler-Marchais, C. D. Klein, M. Bureik, R. Bernhardt, I. Antes, and T. Lengauer. Synthesis and evaluation of imidazolylmethylenetetrahydronaphthalenes and imidazolylmethyleneindanes: Potent inhibitors of aldosterone synthase. *Journal of Medicinal Chemistry*, 48(6):1796–1805, 2005.

[5] M. Voets, I. Antes, C. Scherer, U. Mueller-Vieira, K. Biemel, C. Barrassin, S. Marchais-Oberwinkler, and R. W. Hartmann. Heteroaryl-substituted naphthalenes and structurally modified derivatives: selective inhibitors of CYP11B2 for the treatment of congestive heart failure and myocardial fibrosis. *Journal of Medicinal Chemistry*, 48(21):6632–6642, 2005.

[6] M. Voets, I. Antes, C. Scherer, U. Müller-Vieira, K. Biemel, S. Marchais-Oberwinkler, and R. W. Hartmann. Synthesis and evaluation of heteroaryl substituted dihydronaphthalenes and indenes:

potent and selective inhibitors of aldosterone synthase (CYP11B2) for the treatment of congestive heart failure and myocardial fibrosis. *Journal of Medicinal Chemistry*, 49(7):2222–2231, 2006.

- [7] C. Welsch, F. S. Domingues, S. Susser, I. Antes, C. Hartmann, G. Mayr, A. Schlicker, C. Sarrazin, M. Albrecht, S. Zeuzem, and T. Lengauer. Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of HCV. *Genome Biology*, 9(1):R16, 2008.

29.8.3 Structural Immunoinformatics

Investigators: Kirsten Roomp and Iris Antes

Predicting Epitopes in Large Datasets

A precise understanding of host immune responses is crucial for basic immunological studies as well as for designing effective disease prevention strategies. Epitope-based analysis methods are effective approaches at assessing immune response, allowing for the quantification of the interaction between a host and pathogen, of vaccine effectiveness or other prevention strategies.

Recently, the Immune Epitope Database and Analysis Resource (IEDB) [4] has become available. It is a central data repository and service, containing major histocompatibility complex (MHC) binding data relating to B cell and T cell epitopes from infectious pathogens, experimental pathogens and autoantigens. IEDB is not the first database to store such information, but the first to encompass most of the components and annotation of the previously available databases, as well as the data contained in them.

The experimental screening of large sets of peptides with respect to their MHC binding capabilities is still very demanding due to the large number of possible peptide sequences and the extensive polymorphism of the MHC proteins. Therefore, there is significant interest in the development of computational methods for predicting the binding capability of peptides to MHC molecules, as a first step towards selecting peptides for actual screening.

Our recent DynaPred approach [1] performs a combined structure-sequence-based prediction by incorporating structural information obtained from molecular modeling into a sequence-based prediction model. This method therefore not only allows for the fast prediction of MHC class I binders, but also for the efficient construction of docked peptide conformations. We have evaluated this approach for different MHC class I alleles of the human leukocyte antigen (HLA) genes A and B for which extensive datasets were available in IEDB and compared it to two sequence-based prediction methods from the literature. These two sequence-based prediction methods are the same as examined in Antes et al. [1] and were chosen here as well for comparison reasons. In addition, we have evaluated the prediction server NetMHC which has shown to be among the best predictors in comparison tests [3].

We examined the performance of these four diverse MHC Class I prediction methods on HLA-A and HLA-B allele peptide binding datasets extracted from IEDB [5]. We tested three datasets which differ in the IC50 cutoff criteria used to select the binders and non-binders. The best performance was achieved when predictions were performed on the dataset consisting only of strong binders (IC50 less than 100 nM) and clear non-binders (IC50 greater than 10,000 nM). Robustness of the predictions was only achieved with alleles that were represented with a sufficiently large (greater than 200), balanced set of binders and non-binders. All four methods show good to excellent performance on the comprehensive datasets, with the

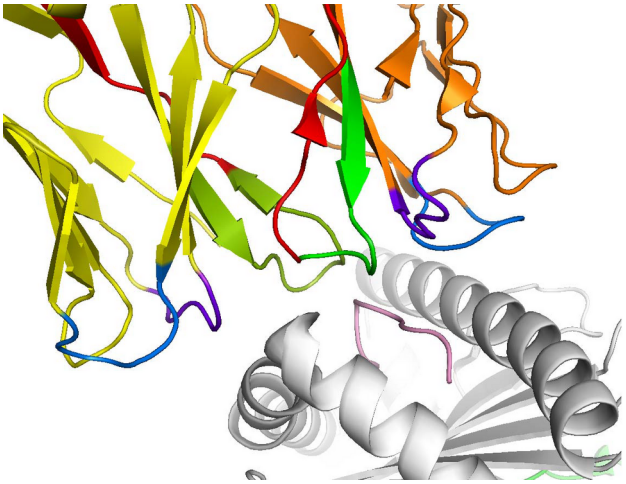


Figure 29.19: The TCR, peptide and MHC complex (PDB: 1OGA). The MHC chain is white, the peptide is pink, the TCRalpha subunit is yellow and the TCRbeta subunit is orange. Complement determining regions (CDRs) of the TCR subunits are colored as follows: CDR1 are blue, CDR2 are purple and CDR3 are green. The J-region of the TCR subunits are colored red. We mutated CDR1, CDR2, CDR3 in both subunits along with the peptide.

artificial neural networks based method outperforming the other methods. However, all methods still struggle to correctly categorize intermediate binders. In a further analysis in which primary anchor residues were excluded, we found that significant information content resides within the non-anchor residues, allowing for reasonable albeit reduced predictive performance even if only those residues are used for training the models.

Simulation Studies with Homology Models of Genetically Engineered CTL TCRs

T-cell receptors (TCRs) are located on the surface of cytotoxic T-cells (CTLs) and recognize antigenic peptides presented by MHCs. This recognition event represents a central event for cellular immune processes and CTLs can then destroy any cell that expresses an appropriate peptide/MHC class I complex.

Antigen-recognition proteins of the specific immune system are complex, having related but highly variable sequences which recognize a huge range of unpredictable ligands. TCRs variability is created during gene rearrangement and is on the order of that of immunoglobulins. The repertoire of TCRs on mature T-cells is tremendously biased toward antigens which consist of peptides bound to MHC. How this bias develops is not yet well understood; so far, the evolutionary selection rules governing the built-in affinity of TCR variable regions have not been evident. A better understanding of this process would impact the design of antigens for vaccination and the choice of optimal TCRs for adoptive T-cell therapy. Research into better understanding these binding rules has also been hampered by the low number of TCR/MHC/peptide complexes for which crystal structures have been solved. Currently, fewer than 30 such complexes are available for analysis, and this comprises both human and

murine complexes.

In collaboration with the Krackhardt Group at GSF in Munich we have developed a strategy for generating homology models of complexes, whose binding affinity has been extensively studied in the laboratory but for which no crystal structures are available [2]. We have selected high-quality human crystal structures, selected the complement determining regions (CDRs), mutated the side chains in the CDRs and built homology models which include appropriate insertions and deletions. Subsequently, the homology models were trimmed to reduce their overall size and used as a starting point for nanosecond-duration molecular dynamics simulations. The simulation results were then compared with both the starting templates and solved crystal structures.

References

[1] I. Antes, S. W.-I. Siu, and T. Lengauer. DynaPred: A structure and sequence based method for the prediction of MHC class i binding peptide sequence and conformations. *Bioinformatics*, 22(14):16–24, 2006.

[2] X. Liang. Structural and functional characterization of the TCR repertoire of Her2369-specific allorestricted T-cells. Master thesis, TU Munich, 2008.

[3] H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz, and V. Brusic. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC Bioinformatics*, 9 Suppl 12:S22, 2008.

[4] B. Peters, J. Sidney, P. Bourne, H.-H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J. V. Ponomarenko, M. Sathiamurthy, S. P. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson, and A. Sette. The design and implementation of the immune epitope database and analysis resource. *Immunogenetics*, 57(5):326–36, 2005.

[5] K. Roomp, I. Antes, and T. Lengauer. Predicting MHC class I epitopes in large datasets. submitted.

29.9 Computational Genomics and Transcriptomics

Coordinator: Thomas Lengauer

29.9.1 Transcriptomics

Investigator: Adrian Alexa

A major challenge of modern genome research is the exploding number of simultaneous genetic measurements for a single patient. For example, with microarray technology the activity of tens of thousand of genes can be measured simultaneously. We develop methods that integrate biological knowledge on gene functions in order to reduce the complexity of explanatory genetic models. Our research in this area focuses on cancer, where we established cooperations with medical and biological experts for prostate cancer and for the brain tumor types meningioma and glioblastoma.

Enrichment Analysis of Gene Ontology Terms

Presently, the research on group testing, also known as gene set enrichment analysis (GSEA), is driven by the field of transcriptomics. Complex biological questions and the continually

growing of biological structured databases require better statistical procedures which integrate more biological (structured) knowledge. Our focus in this field is on better understanding and formalizing the statistical framework of testing the enrichment of nested groups of gene products.

The aim is to identify important biological processes or functions from gene expression data by scoring the relevance of predefined gene categories such as Gene Ontology (GO) categories [5].

Previous Work

We have developed methods that increase the explanatory power of enrichment analysis by accounting for the similarities and dependencies between the gene groups into the calculation of statistical significance [1]. Previously, we introduced two novel methods, namely elim and weight, which provide an approach for finding the most important GO categories on a global scale with respect to graph topology. The algorithms were evaluated both on multiple microarray datasets and on simulated gene expression data.

We investigated alterations of chromosome 8 and hypomethylation of LINE-1 retrotransposons in advanced prostate carcinoma based on a microarray dataset consisting of 24 tumor samples [6, 4]. This study highlighted novel mechanisms in prostate cancer progression and identifies novel candidate genes for diagnostic and therapeutic purposes.

To make our methodology available to the bioinformatic community we developed the topGO software. The package is written in the statistical programming language R and is part of the Bioconductor repository [2]. It is designed to facilitate semi-automated enrichment analysis for Gene Ontology terms. The process (pipeline) consists of normalizing the arrays and inference of gene expression measurements, gene-wise correlation analysis, gene set enrichment analysis, interpretation and visualization of the results. One of the main advantages of topGO, is the unified group testing framework it offers, which enables the user to easily implement new statistical tests or new algorithms that deal with the GO topology. Such an unified framework also facilitates the comparison between different methods or development of new methodology.

Evaluation of Enrichment Analysis

Researchers using Gene Ontology categories for enrichment analysis should understand how gene products are annotated and how the ontologies are structured in order to avoid misinterpretations. We looked into aspects of GO that are usually overlooked and we analyze the consequences of these aspects using simulation studies.

Stability of GO enrichment analysis: A typical enrichment analysis implies testing a set of null hypotheses which should stand true for the GO and the associated annotations. One such null hypothesis states that no GO category should constantly obtain high scores no matter what underling expression dataset is used. However it is not uncommon to see the same GO categories in the results of various enrichment analyses performed on expression data coming from the same microarray chip, although the studies were investigating totally different biological processes. The fact that a set of GO categories always show up as significant shows that there is a bias in the analysis towards these categories.

Performing a stability analysis for the enrichment of GO terms is thus of great interest, though not trivial. The difficulty of this problem, is two-sided. First, one needs to identify the GO categories which under the null hypothesis are significantly enriched and secondly one needs to understand why these GO categories are more highly enriched, on average, than the others. We performed simulation studies on the one hand by permuting genes, on the other by permuting samples to test two different null hypothesis. We have learned that certain GO categories are always reported as significant when the samples are permuted. These categories have constantly high ranks (the most significant GOs) and moreover there is statistical evidence that these categories are deviating from the null hypothesis. By permuting the samples the correlation between genes is preserved. This correlation proved to be crucial for the results in this setup. When the gene correlation is removed, as it is the case with permuting genes, the respective GO categories still obtain high ranks, but there is no statistical evidence to reject the null hypothesis.

We are currently testing to see if the artifacts we have observed are associated with a particular microarray chip or are associated with particular experiments.

Simulation of microarray datasets: Another major challenge in the field of group testing is to provide a framework for comparison and evaluation of the currently available enrichment methods. Some theoretical aspects and the correct biological interpretation of the most used statistical tests are discussed in [3]. The authors divide group testing into competitive and self-contained tests and analyze the advantages of each approach. However, even though the authors try to cover many practical aspects, it is not clear which method would outperform in a particular experimental setup.

We are currently working on novel techniques for generating artificial gene expression datasets, which satisfy the properties of real expression datasets. Advantages of simulation studies include the knowledge on where signals in the data should be found as well as the ability to control the amount of noise that obscures those signals. Nonetheless, simulated data are artificial and barely a weak image of reality. For this reason we developed two novel methods that apply to real expression data sets and at the same time provide the possibility of controlling the location of the true signals.

A first trivial simulation scenario was introduced in [1] and basically defines the genes associated with selected GO categories as significant ones. The second simulation setup takes this idea and generalizes it by permuting gene labels in order to turn the genes associated with selected GO categories into the differentially expressed ones. The GO terms thus selected will be over-represented with respect to the differentially expressed genes. The differentially expressed genes are identified by a by gene-wise analysis performed on a real dataset. In this approach is that the correlation, with respect to the expression measurements, between the genes is preserved, but the association between gene identifiers and GO terms is not. The third simulation study addresses both these issues. It makes use of bi-clustering algorithms to identify a subset of samples and a subset of genes with similar expression patterns within a selected GO category. The expression of the preselected genes is altered such that these genes will exhibit a very strong (most significant) differential expression between the bi-cluster samples and the rest.

The preliminary results showed that methods like weight or elim coupled with statistical

tests like Kolmogorov-Smirnov, two sample t-test, or even Goeman's global test, outperform naive methods which do not account for the topology of the GO hierarchy.

References

[1] A. Alexa, J. Rahnenführer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.

[2] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), 2004.

[3] J. J. Goeman and P. Buhlmann. Analyzing gene expression data in terms of gene sets:methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

- [4] M. Hornstein, M. J. Hoffmann, A. Alexa, M. Yamanaka, M. Müller, V. Jung, J. Rahnenführer, and W. A. Schulz. Protein phosphatase and TRAIL receptor genes as new candidate tumor genes on chromosome 8p in prostate cancer. *Cancer Genomics & Proteomics*, 5(2):123–136, 2008.

- [5] J. Rahnenführer and T. Lengauer. Analysis of expression data: Classification of genes. In T. Lengauer, ed., *Bioinformatics - From Genomes to Therapies 2. Getting at the Inner Workings: Molecular Interactions*, ch. 27, pp. 993–1021. Wiley-VCH, Weinheim, Germany, 2007.

- [6] W. A. Schulz, A. Alexa, V. Jung, C. Hader, M. J. Hoffmann, M. Yamanaka, S. Fritzsche, A. Wlazlinski, M. Müller, T. Lengauer, R. Engers, A. R. Florl, B. Wullich, and J. Rahnenführer. Factor interaction analysis for chromosome 8 and DNA methylation alterations highlights innate immune response suppression and cytoskeletal changes in prostate cancer. *Molecular Cancer*, 6:1–16, 2007.

29.9.2 Analysis of arrayCGH Data

Investigator: Laura Toloşi

Latest developments of microarray technologies reveal new insights on the extent to which DNA copy number variations contribute to genetic variation in both healthy and diseased individuals. Patterns of copy number aberrations have been associated with certain cancer phenotypes and are starting to be used as diagnosis tools in clinics [2]. Recently, high resolution microarrays have been used to show that there exist copy number variations in healthy individuals, referred to as *copy number polymorphisms* (CNPs) [7]. They are a source of human genetic diversity and are thought to play an important role in disease resistance or susceptibility. The research community is making efforts for identifying CNPs in human and characterizing their phenotype.

In spite of the high interest in characterizing the effect of recurrent DNA copy number changes, the need for reliable, automated bioinformatics tools for analysis of copy number alterations is still not met. One of the major reasons is that the latest microarray technologies (arrayCGH, SNP arrays) deliver high-dimensional data to ensure a thorough coverage of the functional parts of the genome. At the same time, due to the high costs of experiments, the number of sampled tissues is often insufficient for carrying out sound statistical analyses.

Our focus is on the discovery of genetic markers relevant for cancer diagnosis and prognosis. We propose a methodology for selection of those copy number aberrations that are most

informative for distinguishing between cancer phenotypes. We adapt statistical learning methods for feature selection to the particular task of quantifying the importance of observed copy number changes for prediction of phenotypes. Previous algorithms have been proposed in the literature, but they either cannot handle high-resolution data [6], or they rely excessively on the assumption that only highly recurrent aberrations are significant [1, 3] although less frequent aberrations may determine particular phenotypes.

Feature Selection Methods for arrayCGH Data

DNA copy number data are typically characterized by a large number of genomic loci at which the log-ratio of abundance of tumor DNA and healthy DNA is experimentally determined. For easier understanding, we refer to these loci as *genes*, although they are often subsequences of genes or other functional parts of the genome.

Say p is the number of genes on the array and n is the number of tumor samples investigated. High resolution experiments yield a very large p and in most of the cases, $p \gg n$. A statistical learning framework can be defined such that the log-ratio of the copy numbers of genes are features, the tumors are samples in the feature space and any phenotype associated to the tumors is a target variable. Ideally, traditional model selection methods can be used to select a classification or regression model that best explains the dependency between the features and the phenotype. Due to the high dimensionality of the feature space, such a solution is statistically and computationally infeasible. Moreover, there is a high correlation between neighboring genes, which needs special handling. Feature selection methods usually underweight correlated features for the sake of model sparsity, while we need to keep all correlated and predictive features, in order to avoid the loss of biologically interesting genomic sites.

We propose to tackle the dimensionality problem by successive steps that restrict the feature space with minimal loss of information. As a first step, we use special smoothing methods [4] to identify breakpoints that delineate regions with different copy number. Apart from a tremendous dimension reduction (the new feature space will yield only as many dimensions as breakpoints found in the tumor collection), this step also ensures more reliable downstream statistical analysis because it factors out the experimental noise from the data. We also propose a second dimension reduction step, which consists of clustering of genes into regions. We rely on the observation that neighboring genes are highly correlated and thus agglomerative hierarchical clustering results in genes grouped together into contiguous regions of the genome. From a different perspective, at this step common breakpoints for all tumor samples are estimated. The final feature set consists of the regions identified by the clustering, ensuring a dimension reduction from several hundred thousand features to only a few hundred.

In most cases, the phenotype is a categorical variable. We use random forests, penalized logistic regression and penalized SVM models for classification. When seeking interpretability, the logistic regression and the random forests are good for the task because the contribution of each feature to the model is quantified. Performance-wise, the random forests and the SVMs are two of the best known instruments for classification.

In order to address the correlated features problem, we bootstrap on the input samples and then average the importance of each feature over all models.

Analysis of Neuroblastoma

We analyzed a cohort of 174 arrayCGH experiments on neuroblastoma tumors with the methods described above. The experiments were performed by our cooperation partners, the group of Prof. Dr. Frank Berthold from the Pediatric Oncology and Hematology Center at Cologne University Clinic. Neuroblastoma is a very interesting case study for our method because the frequency of breakpoints in the respective tumors is significantly lower than in other cancers. It is important for the assessment of the generality of our method to test on cancers that exhibit different copy number aberration patterns. According to its progression status, each tumor in the study was annotated by clinicians with a histopathological grade from I to V, which we use as a phenotype for classification. Results show that low-grade tumors (localized) can be differentiated from high-grade tumors (metastasized) with an accuracy of 86% (SVM, cross-validated) and 84% (random forests). Interestingly, the logistic regression model does not perform as well, leading to the conclusion that there is a non-linear dependency between the copy number changes and the phenotype. We report p-values for each feature to quantify their significance in the random forest models. The list of significant regions is consistent with previous studies on neuroblastoma tumors. Thorough investigation of the genes in the significant regions can lead to the discovery of oncogenes or tumor suppressor genes.

NGFN-Plus Project

We participate in the NGFN-Plus project *Oncogene*, which is funded by BMBF and is a part of the National Genome Research Network. The project is coordinated by Dr. Roman Thomas from Max-Planck-Institute for Neurological Research in Cologne and it involves researchers from the Statistics Department of Dortmund University, MPI of Molecular Physiology in Dortmund, Düsseldorf University, Cologne University and Max Planck Institute for Informatics in Saarbrücken. The goal of this project is to investigate molecular changes in tumors, such as copy number imbalances and genetic mutations and use the gained knowledge to improve cancer therapy. Specialists from different bioinformatic fields will perform wet-lab experiments, data analysis and interpretation. Our task is to create and maintain a database that will afford effective handling of high-dimensional data. We will also design and implement a web-based service that will integrate software for therapy analysis that can assist and improve cancer treatment. We mention that the copy number variation methods presented in this section will be included among the analytic tools of the NGFN-Plus project.

In spite of their high accuracy, non-linear learning models are often used as black boxes, making biological interpretation difficult. We want to estimate oncogenetic models similar to those proposed in [5] in order to gain intuition on the pathways of disease progression, in terms of DNA copy number changes.

References

[1] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, J. Du, T. Kau, R. K. Thomas, K. Shah, H. Soto, S. Perner, J. Prensner, R. M. Debiasi, F. Demichelis, C. Hatton, M. A. Rubin, L. A. Garraway, S. F. Nelson, L. Liau, P. S. Mischel, T. F. Cloughesy, M. Meyerson, T. A. Golub, E. S. Lander, I. K. Mellinghoff, and

W. R. Sellers. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences*, 104(50):20007–20012, 2007.

[2] F. Cappuzzo, F. R. Hirsch, E. Rossi, S. Bartolini, G. L. Ceresoli, L. Bemis, J. Haney, S. Witta, K. Danenberg, I. Domenichini, V. Ludovini, E. Magrini, V. Gregorc, C. Doglioni, A. Sidoni, M. Tonato, W. A. Franklin, L. Crino, J. Bunn, Paul A., and M. Varella-Garcia. Epidermal Growth Factor Receptor Gene and Protein and Gefitinib Sensitivity in Non-Small-Cell Lung Cancer. *J. Natl. Cancer Inst.*, 97(9):643–655, 2005.

[3] S. J. Diskin, T. Eck, J. Greshock, Y. P. Mosse, T. Naylor, C. J. Stoeckert, B. L. Weber, J. M. Maris, and G. R. Grant. STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Research*, 16(9):1149–1158, 2006.

[4] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostat*, 5(4):557–572, 2004.

[5] J. Rahnenführer, N. Beerenwinkel, W. A. Schulz, C. Hartmann, A. von Deimling, B. Wullich, and T. Lengauer. Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics*, 21(10):2438–2446, 2005.

[6] F. Rapaport, E. Barillot, and J.-P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–382, 2008.

[7] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. Large-Scale Copy Number Polymorphism in the Human Genome. *Science*, 305(5683):525–528, 2004.

29.10 System Administration

System administration covers not only maintaining the hard- and basic software (Subsection 29.10.1) for individual group members, but also providing the infrastructure and framework for the large number of web services (Subsection 29.10.2). Indeed in the last to years most of the time was spend for the configuration and performance-tuning of the Oracle XML database for the EpiGRAPH software, the update of our gen2pheno web service and the forum software for the GISAID platform. Last but not least the system administrator introduces new group members and master students into the hard- and software environment of the institute. Consulting therefore plays also a major role in system administration.

Since the two departments "Computational Biology and Applied Algorithmics" and "Computational Genomics and Epidemiology" are working closely together, they share software installations and resources like database, web and compute servers. Systemadministration of department 3 therefore also supports the "Computational Genomics and Epidemiology" group.

29.10.1 Hard- and Software Configuration

Investigator: Joachim Büch

This section informs about the hardware and software resources available to the Department and summarizes the concept for presenting the software offers of the Department to their users.

Software Packages and Bioinformatics Software

Each group member has a Linux desktop and a notebook with dual boot Windows and Linux for his/her personal use. The basic software installation for these machines is carried out by the IST group of the Institute. This group provides the central computer services for all people in the institute. On top of this installation the system administration of the Department provides special software needed for the bioinformatics scientists. We are using scripting languages like Perl, Python, PHP and Ruby as well as statistical software (R, Matlab) and numerous other free available bioinformatics packages.

The software has to be provided not only for 32-bit computer systems, but also for servers with 64-bit architecture, because calculations on very large data volumes need significantly more than 4GB of main memory.

Compute Servers

There are three compute clusters available for the group:

- The group owns a Dell compute cluster with 30 Dell PowerEdge 2650 server nodes with Dual Xeon 3.1 GHz processors and 4 GB RAM per node. The installed Message Passing Interface facilitates running parallel programs on the nodes. Projects working with Gromacs software and DynaCell, which was developed in the group (Section 29.8.1), are the primary users of this cluster. By now, the hardware is outdated and we plan to replace the 30 nodes by 8 newer Quad Core servers in near future.
- The Institute provides a Linux compute cluster with 100 nodes, where each node is a Suns V20z with the 64-bit architecture of two Dual Core Opterons with 8 GB memory. Job queuing is facilitated via Sun's "N1 Grid Engine" software. Jobs on this cluster can only be used in batch mode.
- Additional six powerful compute servers with 4 x Quad Core Xeon or Opteron 2,93 GHz processors and 128 GB RAM are available, on which programs can be run interactively. This is needed for rapid prototyping of programs written with scripting languages. Being only used by the bioinformatics groups, these servers can easily configured with new software according to the needs of the group.

Database Servers

Several database servers are available. The group is using MySQL databases on Debian Linux platforms and Oracle databases on Solaris 10 operating systems.

An older database server is still running MySQL software, version 4.1, on a Dual Core Xenon CPU with 270 GB local disk space available. Internet content information, displayed on the groups homepages and technical information about the webservice configuration is stored on this database Most bioinformatics data are hosted in a MySQL 5.0 database on a Sun V40z with two Dual Core 2.5 GHz processors and 16 GB RAM. This database is also used by web services running on the public internet server. Disk space for the database is provided by an external disc array. In addition there are a few more MySQL databases located on webservers. We use only 64-bit versions of MySQL.

Oracle database servers are used for software with a huge amount of data or if XML data has to be be accessed efficiently. Since 64-bit Oracle software is not available for Debian Linux, we installed our Oracle servers on top of Solaris 10 operating system. Two servers are available: Oracle 10g on a machine with 4 x Quad Core Opteron 2.8GHz with 64 GB RAM and Oracle 11g on a SPARC Enterprise T5220 with 64 1.4GHz UltraSPARC-T2 processors and 64 GB RAM. Here again disk space of 4TB is provided by a Storage Area Network (SAN).

Web Servers

The majority of the web services and the intranet content management system is hosted on a three-year old Sun V40z web server with two 2.5MHz Opteron processors. Load balancing is done by sending requests for larger computations to another dedicated 4 x Dual Core Opteron 2.8GHz processor server with 32 GB RAM. Several Tomcat and RPC services are running on that machine, communicating with the Apache server on the web server.

The hardware for a new web server, a Dell PowerEdge with 4 x Quad Core Xeon 2.9 processors and 64 GB RAM is already available, but not yet installed.

The GISAID platform is located on its own webserver, a Sun Fire X4100 with 2 x Dual Core Opteron and 8 GB RAM.

Storage Area Network (SAN)

Two independent disk arrays with 3.6TB each are providing slices of a Raid 5 disk storage for the MySQL, Subversion and Oracle servers. By means of an attached Fibre Channel switch, located in-between the database servers and the SAN, the available disk space can dynamically switched among the database servers.

Future Hardware Plans

Since our Oracle and MySQL databases are still growing rapidly, we need two more SAN disk arrays, that can be attached to the Fibre Channel switch. In addition the old PowerEdge Cluster needs to be replaced.

29.10.2 Infrastructure for Web Services

Investigator: Joachim Büch

As already mentioned before, providing an appropriate technical framework for web services is the most important task, the system administration has to perform. More and more publications in the group come along with a more or less complicated web service, offering a bioinformatics analysis to the scientific community. Up-to-date databases, a toolbox for automating the process of creating a new web service and concepts for load balancing are needed to fulfill these requirements.

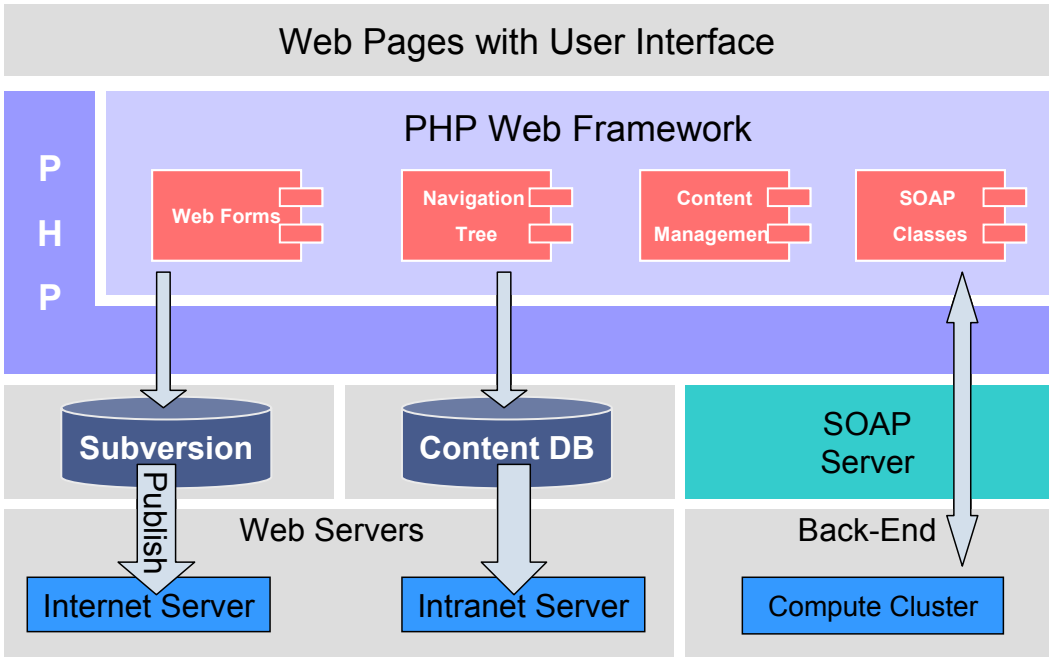


Figure 29.20: The web framework automates the creation of new websites.

PHP Web Framework and Intranet Corporate Design

We are still refining and improving our PHP Web Framework, which the group is using as basis for new web and intranet services. We started developing such a system several years ago, when we were conceptualizing a technique allowing for separating the web front-end from the program logic. It was our intention to have basic templates for the websites, automatically providing the Max Planck Institutes' corporate identity. Moreover the websites should be independent from the web server hardware, the operating system and the base URL of the project, thus allowing for running the same web pages on different web servers. We started developing a PHP-based object-oriented class library collecting all configuration data and basic information for the website in a database (see Figure 29.20). Web sites can now be developed and tested on an internal web server and the revision control system Subversion is used to keep track of changes in the webpages. Common tasks like accepting license agreements, downloading open source software and presenting the results of the analysis have been automated. The framework is also used for providing information about installed software and hardware, seminar talks and so on in our intranet. Some work still has to be done to completely automate the publishing process of web services. Each group member should be able to update and configure his/her own web services without the help of the system administrator, and such that other services are not affected.

Available Bioinformatics Web Services

We have several web services, offering downloads of bioinformatics software, that was developed in the group. In such cases users will find an online documentation and a download page, where they can register and download the software, that can be installed locally on their desktop. However most of our new web services need a large amount of data for predicting results. Huge data models for Protein Function Predictions (see the GOdot system in Subsection 29.7.2), FunSimMat, a functional similarity database for proteins (see Subsection 29.5.3), HIV Drug Resistance prediction, HIV therapy optimization (see Subsection 29.4.1) and the Computational Epigenetics project (see EpiGRAPH [5] software in Subsection 29.6.1) are kept in Oracle and MySQL databases. We do not offer direct access from the Internet to those databases, but instead offer online web services with comfortable graphical user interfaces to the data stored in the databases. Such services cannot be distributed as stand-alone software tools together with data models.

The following table shows a complete list of all available web services. The web interfaces for the web services were developed with the PHP framework described in the section above. This facilitates a strict separation between the web design with the look and feel of the website and the program logic behind (MVC Model View Controller concept):

Arby [7]	http://arby.bioinf.mpi-inf.mpg.de
BiLayout	http://bilayout.bioinf.mpi-inf.mpg.de
BioMyn 29.5.1	http://www.biomyn.de
BiQ Analyzer 29.6	http://biq-analyzer.bioinf.mpi-inf.mpg.de
CalSpec	http://bioinf.mpi-sb.mpg.de/projects/CalSpec
CgiHunter	http://cgihunter.bioinf.mpi-inf.mpg.de
DASMI 29.5.1 [4]	http://dasmi.de
DomainGraph 29.5.2	http://domaingraph.bioinf.mpi-inf.mpg.de
DomainNetworkBuilder	http://med.bioinf.mpi-inf.mpg.de/domainnet
EpiGRAPH 29.6.1	http://epigraph.mpi-inf.mpg.de
Euresist 29.4.1	http://engine.euresist.org
FunSimMat 29.7.2	http://funsimmat.bioinf.mpi-inf.mpg.de
Geno2Pheno Coreceptor 29.4.2	http://coreceptor.bioinf.mpi-inf.mpg.de
Geno2Pheno Integrase 29.4.1	http://integrase.bioinf.mpi-inf.mpg.de
Geno2Pheno Resistance 29.4.1	http://www.geno2pheno.org
Galinter 29.7.3	http://galinter.bioinf.mpi-inf.mpg.de
Genafor 29.4	http://www.genafor.org
GOdot 29.7.2	http://godot.bioinf.mpi-inf.mpg.de
GOTax 3	http://www.gotaxexplorer.de
HBV Drug Resistance	http://www.genafor.org/hbv/hbvpredict.php
IRECS 29.8.1	http://irecs.bioinf.mpi-inf.mpg.de
methmarker 29.6	http://methmarker.mpi-inf.mpg.de
Mtreemix 3	http://mtreemix.bioinf.mpi-inf.mpg.de
NetworkAnalyzer	http://med.bioinf.mpi-inf.mpg.de/netanalyzer
NOXclass 29.7.3	http://noxclass.bioinf.mpi-inf.mpg.de
Recco	http://recco.bioinf.mpi-inf.mpg.de

ROCR [7]	`http://rocr.bioinf.mpi-sb.mpg.de`
STRuster 29.7.1	`http://struster.bioinf.mpi-inf.mpg.de`
topGo 29.9.1	`http://topgo.bioinf.mpi-inf.mpg.de`
TSMap	`http://tsmap.bioinf.mpi-inf.mpg.de`
ViralDAS	`http://viraldas.bioinf.mpi-inf.mpg.de`

Calculations for web services with a short response time run directly on the internet server and the result is presented immediately on the website. Services that need a longer compute time are working asynchronously. The web front-end on the internet server communicates by means of a SOAP protocol with the back-end service on the compute server or compute cluster behind the institutes firewall. After completion, the back-end generates a dynamic website with the results and sends an email with the URL of this site to the user. Web services requiring more complex input by the user provide a small downloadable client program to the user. Running on the user's desktop, this program sends data over the Internet to our web service and presents the result of the prediction after completion to the remote user.

Typically, together with the researchers, the system administrator chooses the best fitting software concept for the needs of the individual projects and helps designing the database.

geno2pheno

The geno2pheno software package provides a rule based prediction tool for HBV drug resistance on basis of the viral genotype geno2pheno[hbv] and four prediction services for HIV drug resistance based on mutation lists of the DNA of the virus:

geno2pheno[resistance] predicts drug resistance against reverse transcriptase and protease inhibitors, geno2pheno[theo] [1] predicting response to antiretroviral therapy on basis of the viral genotype and the drugs in the regimen, geno2pheno[coreceptor] [9] predicts the coreceptor usage of the virus and geno2pheno[integrase] is an interpretation tool for resistance mutations against the new class of integrase inhibitor.

The services are used by virologists in their daily work for treatment of HIV and HBV infected people. The user interface of these web services was therefore adapted to their needs and the old existing geno2pheno[resistance] service was completely redesigned. Four of these services are now using the PHP toolbox – mentioned above – thus presenting the same look and feel. For performance reasons, parts based on scripting language in the old geno2pheno[resistance] server have been replaced with native program code in C++ step by step and the resulting machine code was directly linked into the Apache web server as loadable module. In addition we are constantly updating the service, such as if genotype-phenotype training data become available for new drug on the marketplace.

GISAID

The GISAID Platform (Global Initiative for Sharing Avian Influenza Data, `http://www.gisaid.org`) is an independent multinational initiative that offers sequence and tracking data on influenza strains. The platform comprises three parts:

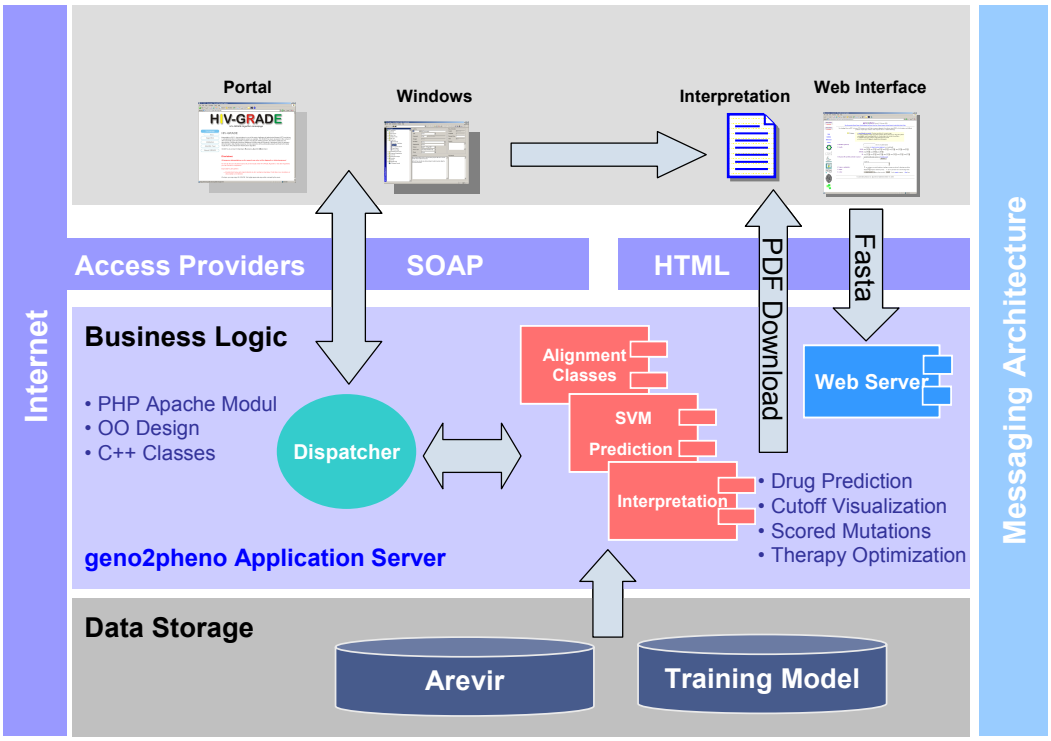


Figure 29.21: The new software design of GENO2PHENO uses the SOAP protocol to link the GUI with the server-side back-end, thus allowing to use different clients for GENO2PHENO. For the sake of a better performance, all program objects are linked together in an Apache module.

A portal software with wiki, closed forums and user administration, the Epiflu sequence database (hosted by the Swiss Institute for Bioinformatics, Lausanne) and a tracking system for the virus samples (hosted by Kisters AG, Duisburg). The software for these systems is hosted on different places.

The Department is hosting a number of information, registration and networking offers in a content management and provides a single-logon mechanism for the EpiFlu database and the tracking system. We used the commercial software dante®portal, a component based framework developed by a3systems, Saarbrücken. This system facilitates integration of the EpiFlu database and the tracking system and development of additional modules, such as a complex workflow for the registration process, news feed integration and a special member area for registered users.

High Availability of Web services

Our web and database servers are not mirrored. Nevertheless we have prepared an infrastructure, that allows for replacing any hardware with a minimum downtime for the web services:

Database files are located on Raid5 Storage Area Networks (SAN), working with Raid level 5 disk arrays. For safety reasons we have two independent devices mirrored by the operating system of the database server. Database servers can simply be replaced by switching the SAN to another machine. Using Solaris 10 with ZFS as filesystem, it is possible to make cold backups of even huge databases within seconds by means of ZFS snapshots. Using additionally Oracle archive files prevents any data loss.

Configuration files for the webserver can be generated automatically by means of the technical information about the web services, stored in our intranet database. Finally the programs for the web service, being kept in a source code revision control system, can be restored by an automatic checkout.

Databases

Some bioinformatics databases come as flat files, but the majority of projects now use MySQL or Oracle databases for performance reasons. Specifically the EpiGRAPGH project (see Subsection 29.6.1), which currently stores 500 GB of data in the database, benefits from Oracle's XML DB feature and the fast SQL query processing. During the whole development process a permanent tuning of the database design had to be done to reach the best performance. Especially the performance tuning of the new XML Queries was a challenging and tedious process and finally with the Oracle 11g version, we succeeded in speeding up queries by a factor of 10.

For the geno2pheno project see Subsection 29.4.1), we are hosting a copy of the *Arevir* database [8] containing viral DNA sequences for the analysis of resistance mutations of the human immunodeficiency virus.

Source Code Revision Control System

The Revision Control System Subversion is used heavily by the group. All software developed in the group and all web services are archived in Subversion. Our framework for the web services integrates the publishing process of websites in Subversion, thus ensuring an appropriate level of quality management for the public web services and project software.

Conferences

Since 2006 we are providing the internet representation of the GENAFOR society with the website `http://www.genafor.org/events.php` for the annual Arevir meetings in Bonn. The geno2pheno HIV web services as well as the HBV prediction service are identically linked in the sitemap of the GENAFOR 29.4 homepage.

References

- [1] A. Altmann, M. Däumer, N. Beerenwinkel, Y. Peres, E. Schülter, J. Büch, S.-Y. Rhee, A. Sönnerborg, W. J. Fessel, R. W. Shafer, M. Zazzi, R. Kaiser, and T. Lengauer. Predicting the response to combination antiretroviral therapy: Retrospective validation of geno2pheno-THEO on a large clinical database. *The Journal of Infectious Diseases*, 199:999–1006, 2009.

- [2] A. Altmann, M. Rosen-Zvi, M. Prosperi, E. Aharoni, H. Neuvirth, E. Schülter, J. Büch, Y. Peres, I. Incardona, A. Sönnerborg, R. Kaiser, M. Zazzi, and T. Lengauer. The EuResist approach for predicting response to anti HIV-1 therapy. In *Reviews in Antiviral Therapy*, Utrecht, Netherlands, 2008, vol. 2008, p. 107. Virology Education.
- [3] A. Altmann, M. Rosen-Zvi, M. Prosperi, E. Aharoni, H. Neuvirth, A. Sönnerborg, E. Schülter, J. Büch, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi, and T. Lengauer. Comparison of classifier fusion methods for predicting response to anti HIV-1 therapy. *PLoS ONE*, 3(10):e3470, 2008.
- [4] H. Blankenburg, R. D. Finn, A. Prlić, A. M. Jenkinson, F. Ramírez, D. Emig, S.-E. Schelhorn, J. Büch, T. Lengauer, and M. Albrecht. DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10):1321–1328, 2009.
- [5] C. Bock, K. Halachev, J. Büch, and T. Lengauer. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. *Genome Biology*, 10:R14, 2009.
- [6] L. Feuerbach, B. Christopher, K. Halachev, J. Büch, and T. Lengauer. Towards comparative epigenomics: A software toolkit for cross-species epigenome data analysis. In G. Meijer, T. Ried, W. Giaretti, and P. Hamilton, eds., *Abstracts of the 2nd MC-GARD Meeting*, Madrid, Spain, 2008, *CELLULAR ONCOLOGY Special Issue*, vol. 30, p. 230. IOS.

[7] M.-P.-I. für Informatik. Seventh biennial report. Seventh Biennial Report, 2005.

[8] K. Roomp, N. Beerenwinkel, T. Sing, E. Schülter, J. Büch, S. Sierra-Aragon, M. Däumer, D. Hoffmann, R. Kaiser, T. Lengauer, and J. Selbig. Arevir: A secure platform for designing personalized antiretroviral therapies against HIV. *Lecture Notes in Computer Science*, 4075:185–194, 2006.

- [9] T. Sing, A. J. Low, N. Beerenwinkel, O. Sander, P. K. Cheung, F. S. Domingues, J. Büch, M. Däumer, R. Kaiser, T. Lengauer, and P. R. Harrigan. Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antiviral Therapy*, 12(7):1097–1106, 2007.

29.11 Academic Activities

29.11.1 Journal Positions

Thomas Lengauer is a member of the editorial boards of:

- *Bioinformatics (Associate Editor)* (since 1996),
- *IEEE-ACM Transactions on Computational Biology and Bioinformatics* (since 2004),
- *Journal of Computational Biology* (since 1997),
- *PLoS Computational Biology* (since 2007),
- *Springer Lecture Notes in Bioinformatics* (since 2003).

29.11.2 Conference and Workshop Positions

Membership in Program Committees

Thomas Lengauer:

- *13th Annual International Conference on Computational Molecular Biology (RECOMB 2009)*, Tuscon, Arizona, USA, May 2009

- *7th European Conference on Computational Biology (ECCB 2008)*, Cagliari, Sardinia, September 2008
- *12th Annual International Conference on Computational Molecular Biology (RECOMB 2008)*, Singapore, March/April 2008

Adrian Alexa:

- *16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2008)*, Toronto, Canada, July 2008

Hiroto Saigo:

- *1st Asian Conference on Machine Learning*, Nanjing, China, November 2009

Membership in Organizing Committees

Thomas Lengauer:

- *13th Annual International Conference on Computational Molecular Biology (RECOMB 2009)*, Tuscon, Arizona, USA, May 2009
- *7th European Conference on Computational Biology (ECCB 2008)*, Cagliari, Sardinia, September 2008
- *16th International Conference on Intelligent Systems for Molecular Biology (ISMB 2008)*, Toronto, Canada, July 2008
- *Recent Challenges for Statistics in the Biosciences - 100 Years after Gustav Zeuner*, Freiberg, Saxonia, January 2008 (with with Dietrich Stoyan, Freiberg, Niels Keiding, Peter Fritz, Leipzig, Hans Föllmer, Berlin)
- *15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2007) & 6th European Conference on Computational Biology (ECCB 2007)*, Vienna, Austria, July 2007
- *Arevir Workshop on HIV Resistance Analysis*, Bonn, Germany, April 2007 (with Rolf Kaiser, Cologne)

Mario Albrecht:

- *15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2007) & 6th European Conference on Computational Biology (ECCB 2007)*, Vienna, Austria, July 2007

29.11.3 Invited Talks and Tutorials

Thomas Lengauer:

- *Talks on HIV Bioinformatics*
  - Symposium "Systems Biology and Translational Pneumology", Mannheim, March 2009
  - University of Leuven, February 2009

    - University of Freiburg, February 2009
    - University of Konstanz, February 2009
    - University of Hamburg, January 2009
    - Workshop on Machine Learning, NIPS 2008, Vancouver, Canada, December 2008
    - 4th HIV Entry Workshop, Puerto Rico, November 2008
    - University of Heidelberg, October 2008
    - Annual Meeting of the German Society for Fighting Infectious Diseases (DVV), Cologne, September 2008
    - Helmholtz High School, Bonn, August 2008
    - Public Lecture, Annual Meeting of Max Planck Society, Dresden, June 2008
    - European Patent Office, April 2008
    - Bioperspectives 2007, Cologne, May 2007
- *Epigenetics*, Helmholtz High School, Bonn, March 2008
- *Protein Structure, Function and Interactions - Basic Research and Applications*, 3DSig at ISMB/ECCB 2007, Vienna, July 2007
- *Bioinformatics - Computer-based Search for New Drugs*, University of Karlsruhe , June 2007

Mario Albrecht:

- *Network Bioinformatics for Understanding Protein Interactions*, Invited talk, Seminar Series of the Graduate School of Biotechnology & Biochemistry & Molecular Sciences, University of Padova, Padova, Italy, February 2009.
- *Using DASMI to Exchange and Annotate Molecular Interactions*, Invited talk, Tutorial on Interoperability of Bioinformatics Software and Databases at the 7th European Conference on Computational Biology (ECCB), Cagliari, Italy, September 2008.
- *Bioinformatics Tools for Systems Biology Work*, Invited talk, Invited Talk Series, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands, September 2008.
- *Overview on Biological Network Visualizations*, Invited talk, Seminar on Graph Drawing with Applications to Bioinformatics and Social Sciences, Schloss Dagstuhl, Wadern, Germany, May 2008.
- *Investigating Domain and Residue Interaction Networks of Proteins for Gaining Biomedical Insight*, Invited talk, Basel Bioinformatics Seminar, Biozentrum, University of Basel, Basel, Switzerland, April 2008.
- *Computing Topological Parameters with NetworkAnalyzer*, Invited talk, 5th Annual Cytoscape Symposium, Amsterdam, The Netherlands, November 2007.
- *Using Cytoscape Tools to Evaluate Protein Interaction Data*, Invited talk, AAPS National Biotechnology Conference, San Diego, CA, USA, June 2007.
- *Molecular Networks in Medical Bioinformatics: Understanding Human Cells*, Invited talk, IOI Seminar at the International Conference und Research Center for Computer Science (IBFI), Schloss Dagstuhl, Wadern, Germany, May 2007.

- *Quality Evaluation of Human Protein Networks*, Invited talk, HUPO Proteomics Standards Initiative (PSI) Spring Meeting, Lyon, France, April 2007.

Adrian Alexa:

- *Scoring Gene Ontology Terms* , Tutorial, NGFN Workshop: Courses in Practical DNA Microarray Analysis, Munich, Germany, May 2008.
- *Scoring Gene Ontology Terms* , Tutorial, NGFN Workshop: Courses in Practical DNA Microarray Analysis, Munich, Germany, November 2007.
- *Scoring Gene Ontology Terms* , Tutorial, NGFN Workshop: Courses in Practical DNA Microarray Analysis, Dortmund, Germany, Septemebr 2007.
- *Scoring Gene Ontology Terms* , Tutorial, NGFN Workshop: Courses in Practical DNA Microarray Analysis, Heidelberg, Germany, March 2007.

André Altmann:

- *Do Not Blame the Fitness! Only Slight Impact of Predicted Replicative Capacity for Therapy Response Prediction*, Invited talk, 7th European HIV Drug Resistance Workshop, Stockholm, Sweden, March 2009.
- *Keeping Models that Predict Response to Antiretroviral Therapy Up-to-date: Fusion of Pure Data-driven Approaches with Rules-based Methods*, Invited talk, 7th European HIV Drug Resistance Workshop, Stockholm, Sweden, March 2009.
- *Inferring Virological Response from Genotype: With or Without Predicted Phenotypes?*, Invited talk, Virco, Mechelen, Belgium, December 2008.
- *Bioinformatics: RESINA and geno2pheno*, Invited talk, 2nd HIV-Resistenz-Workshop, Robert Koch Institut Berlin, Germany, November 2008.
- *Interpretation and Visualization of HIV Drug Resistance in Ultra-Deep Sequenced Virus Populations*, Invited talk, 1st European Meeting on 454 Sequencing of Virus Populations, Penzberg, Germany, October 2008.
- *Ultra-deep Sequencing Data from the Bioinformatics Point of View: Challenges and Requirements for the Analysis*, Invited talk, 1st European Meeting on 454 Sequencing of Virus Populations, Penzberg, Germany, October 2008.
- *Prediction: Classification and Regression*, Tutorial, 14th International BioInformatics Workshop on Virus Evolution and Molecular Epidemiology, South African National Bioinformatics Institute Cape Town, South Africa, September 2008.
- *Guiding Treatment Selection: Past, Present, and Future*, Invited talk, Arevir Meeting 2008, Bonn, Germany, April 2008.
- *geno2pheno[integrase]*, Invited talk, Arevir Meeting 2008, Bonn, Germany, April 2008.
- *The EuResist Approach for Predicting Response to Anti HIV-1 Therapy*, Invited talk, 6th European HIV Drug Resistance Workshop, Budapest, Hungary, March 2008.
- *Prediction of Response to Combination Antiretroviral Therapy for HIV Infection*, Invited talk, Fraunhofer Institute, Kaiserslautern, Germany, August 2007.

- *Prediction: Classification and Regression*, Tutorial, 13th International BioInformatics Workshop on Virus Evolution and Molecular Epidemiology, Instituto Nacional de Saude Dr. Ricardo Jorge (INSA), Lisbon, Portugal, September 2007.
- *Clinical Relevance of Quantitative Resistance Information*, Invited talk, Deutsch-Österreichischer AIDS Kongress 2007, Frankfurt, Germany, June 2007.
- *geno2pheno-THEO - Therapy Optimizer*, Invited talk, Arevir Meeting 2007, Bonn, Germany, April 2007.
- *geno2pheno-THEO: Predicting Clinical Outcome of Combination ART*, Invited talk, Arevir Meeting 2007, Bonn, Germany, April 2007.
- *Validation of geno2pheno and THEO on a Large Independent Clinical Dataset*, Invited talk, 5th European HIV Drug Resistance Workshop, Cascais, Portugal, March 2007.
- *Improved Prediction of Response to Antiretroviral Combination Therapy using the Genetic Barrier to Drug Resistance*, Invited talk, Statistik 2007, Bielefeld, Germany, March 2007.

Iris Antes:

- *Protein-ligand Docking Including Protein Flexibility - A Hierarchical Approach* , Invited talk, From Computational Biophysics to Systems Biology Workshop, Jülich, Germany, May 2008.
- *Kombinierte biophysikalisch-bioinformatische Methoden für Wirkstoff-design und Proteinmodellierung*, Invited talk, Pharmaceutical Seminar, University of Tübingen, Germany, May 2008.
- *Towards Biophysical Accuracy in Drug Design and Immunoinformatics*, Invited talk, Workshop on Modern Trends in Scientific Computing, University of Lugano, Switzerland, April 2008.
- *Kombinierte biophysikalisch-bioinformatische Methoden zum Wirkstoff-design und zur Proteinmodellierung*, Invited talk, Chemistry Seminar, University of Graz, Germany, April 2008.
- *DynaDock: Protein-Ligand Docking Including Protein Flexibility*, Invited talk, Workshop on Computer Simulation and Theory of Macromolecules, Hünfeld (Germany), April 2008.
- *Structure-based Prediction of MHC-peptide Binding and HIV-drug Resistance*, Invited talk, Immunology Seminar, Technical University of Munich and GSF, Germany, December 2007.
- *Kombinierte biophysikalisch-bioinformatische Methoden zum Wirkstoff-design und zur Proteinmodellierung*, Invited talk, Pharmaceutical Chemistry Seminar, University of Düsseldorf, Germany, October 2007.
- *Combined Bioinformatics and Biophysical Methods for Molecular Docking and Protein Modeling*, Invited talk, Computer Science Seminar, Bergen, Norway, October 2007.
- *Combined Bioinformatics and Biophysical Methods for Prediction in Chemical Biology*, Invited talk, Indo-German Conference: Modeling Chemical and Biological (Re)acticity, Hyderabad, India, September 2007.

- *Predicting Bound Protein-peptide Conformations: Application to MHC-peptide Complexes*, Invited talk, 234th National Meeting of the American Chemical Society, Boston, USA, August 2007.
- *A New Approach for Flexible-ligand Flexible-receptor Docking*, Invited talk, 3DSig Workshop at the 15th Annual International Conference on Intelligent Systems in Molecular Biology (ISMB) 2007, Vienna, Austria, July 2007.
- *Kombinierte biophysikalisch-bioinformatische Methoden zur Wirkstoff- und Proteinmodellierung*, Invited talk, Biological Chemistry Seminar, Technical University of Munich, Germany, July 2007.

Hagen Blankenburg:

- *DAS for Molecular Interactions*, Invited talk, DAS Workshop, Hinxton, United Kingdom, March 2009.
- *The PSISCORE Molecular Interaction Confidence Scoring Framework*, Invited talk, HUPO Proteomics Standards Initiative (PSI) Workshop: Development of standards-compliant tools for molecular interaction data management, Hinxton, United Kingdom, November 2008.
- *PSI Common Confidence Scoring System*, Invited talk, HUPO Proteomics Standards Initiative (PSI) Spring Meeting, Toledo, Spain, April 2008.
- *DAS for Molecular Interactions*, Invited talk, DAS Workshop, Hinxton, United Kingdom, February 2008.

Christoph Bock:

- *Computational Epigenetics Applied to Cancer Research*, Invited talk, Dana Farber Cancer Institute, Cambridge, MA, February 2009.
- *Biomedical Research Seminars: Epigenome Analysis and Cancer*, Invited talk, Université Libre de Bruxelles, Brussels, Belgium, January 2009.
- *Bioinformatic Methods for Epigenetic Biomarker Discovery*, Invited talk, CHI Workshow: Advances in DNA Methylation Analysis, Boston, MA, October 2008.
- *Bioinformatic Methods for Epigenome Analysis*, Invited talk, Pfizer Regenerative Medicine, Sandwich, UK, August 2008.
- *Epigenome Analysis by Statistical Learning Methods*, Invited talk, CAS-MPG Partner Institute for Computational Biology, Shanghai, China, July 2008.
- *Machine Learning Helps Pinpoint Epigenetic Damage in Cancer*, Invited talk, University of Lübeck Medical School, Lübeck, Germany, June 2008.
- *Powerful Bioinformatic Methods for Cancer Epigenomics*, Invited talk, Keystone Symposium: Cancer Genomics and Epigenomics, Taos, NM, February 2008.
- *Machine Learning in Computational Epigenetics*, Invited talk, Friedrich Miescher Laboratory, Tübingen, Germany, January 2008.
- *EpiGRAPH: Prediction as a Step Toward Understanding of Epigenomes*, Invited talk, CSHL/WT Genome Informatics Meeting, Cold Spring Harbor, NY, November 2007.

- *Bioinformatics series: Computational Epigenetics*, Invited talk, Broad Institute of MIT and Harvard, Cambridge, MA, November 2007.
- *Realizing the Medical Potential of Human Epigenome Mapping*, Invited talk, ESF Exploratory Workshop on Computational Approaches to the Role of Epigenetic Marks in Transcription Regulation, Basel, Switzerland, October 2007.
- *Strategies for Efficient DNA Methylation Mapping and Data Analysis*, Invited talk, EU Network of Excellence: The Epigenome - Second Epigenome Mapping Workshop, Berlin, Germany, September 2007.
- *Realizing the Medical Potential of Human Epigenome Mapping*, Invited talk, ISMB/ECCB 2007 Special Session on Computational Epigenetics, Vienna, Austria, July 2007.

Dorothea Emig:

- *Visualizing the Analysis of Interaction Networks in Combination with Alternative Splicing*, Invited talk, University of California, Santa Cruz, CA, USA, March 2009.
- *Visualizing Interaction Networks and the Effects of Alternative Splicing*, Invited talk, Dortmund University of Technology, Dortmund, Germany, December 2008.
- *Integrative Visual Analysis of the Effects of Alternative Splicing on Protein Domain Interaction Networks*, Invited talk, 5th Annual Meeting of the International Symposium on Integrative Bioinformatics (IB08), Lutherstadt Wittenberg, Germany, August 2008.
- *Integration of Expression Data with Interaction Networks Using the DomainGraph Plugin*, Invited talk, 6th Annual Cytoscape Retreat, Toronto, Canada, July 2008.
- *Visualizing Domain Interaction Networks and the Impact of Alternative Splicing Events*, Invited talk, 2nd International Symposium of Information Visualization in Biomedical Informatics (IVBi), London, United Kingdom, July 2008.
- *Visualizing Domain Interaction Networks and the Impact of Alternative Splicing Events*, Invited talk, Cancer Research UK, London, United Kingdom, July 2008.
- *Cytoscape Application Showcases*, Tutorial, 9th International Conference in Systems Biology, Gothenburg, Sweden, 2008.
- *Introduction to the BiLayout and DomainGraph Plugins*, Invited talk, 5th Annual Cytoscape Retreat, Amsterdam, Netherlands, November 2007.

Lars Feuerbach

- *Comparative Epigenomics: Cross-species analysis of CpG islands*, Invited talk, Genetics/Epigenetics Department, Saarland University, Saarbrücken, Germany, January 2009.

Fidel Ramírez:

- *Data integration for the life sciences*, Invited talk, Genetics/Epigenetics Department, Saarland University, Saarbrücken, Germany, December 2008.

Hiroto Saigo:

- *Iterative Subgraph Mining for Principal Component Analysis*, Invited talk, 8th IEEE International Conference on Data Mining (ICDM2008), Pisa, Italy, December 2008.

- *Partial Least Squares Regression for Graph Mining*, Invited talk, 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2008), Las Vegas, USA, August 2008.

Oliver Sander:

- *GODOT: Inferring Molecular Function from Local Function Conservation in Sequence and Structure Space*, Invited talk, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan, January 2008.

Sven-Eric Schelhorn:

- *An Integrative Approach for Predicting Interactions of Protein Regions*, Invited talk, 7th European Conference on Computational Biology (ECCB), Cagliari, Italy, September 2008.
- *Integrative Prediction and Analysis of Protein Region Interactions*, Invited talk, 4th CSHL Meeting on Network Biology, Hinxton, United Kingdom, August 2008.

Andreas Schlicker:

- *Applications of Semantic Similarity Measures*, Invited talk, Ontologies and Text Mining for Life Sciences: Current Status and Future Perspectives, Schloss Dagstuhl, Wadern, Germany, March 2008.
- *FunSimMat: A Comprehensive Functional Similarity Database*, Invited talk, NGFN SMP Bioinformatics workshop "From Bioinformatics to Medical Systems Biology", German Cancer Research Center (DKFZ), Heidelberg, Germany, November 2007.

Ingolf Sommer:

- *Structural Bioinformatics at the Max-Planck-Institute for Informatics*, Invited talk, Journee Bioinfo at LORIA/INRIA, Nancy, France, September 2007.
- *Moment Invariants as Shape Recognition Technique for Comparing Protein Binding Sites*, Invited talk, ISMB 3D-SIG, Vienna, Austria, July 2007.
- *From Protein Structure to Function*, Invited talk, Bio-Mathematics and Informatics Symposium, CWI, Amsterdam, The Netherlands, April 2007.

Alexander Thielen:

- *Performance of Genotypic Coreceptor Measurement Using geno2pheno[coreceptor] in B- and non-B HIV Subtypes in a Large Cohort of Therapy-experienced Patients in Germany*, Invited talk, 7th European HIV Drug Resistance workshop, Stockholm, Sweden, March 2009.
- *Improved Genotypic Prediction of HIV-1 Coreceptor Usage by Incorporating V2 Loop Sequence Variation*, Invited talk, XVII International HIV Drug Resistance Workshop: Basic Principles and Clinical Implications, Sitges, Spain, June 2008.
- *Genotypic Analysis of Coreceptor Usage - New Developments and Applications for geno2pheno[coreceptor]*, Invited talk, Arevir Meeting 2008, Bonn, Germany, April 2008.

- *Bestimmung von R5/X4 Virus im Rahmen der ART*, Invited talk, Münchner AIDS Tage, Berlin, Germany, March 2008.
- *Tropismustestung in Deutschland - Datenübersicht und Vergleich: Genotyp vs. Phänotyp*, Invited talk, Pfizer, Frankfurt, Germany, February 2008.
- *Improving HIV-1 Coreceptor Usage Prediction from Genotype with Regions Outside of V3*, Invited talk, 3rd International Workshop on Targeting HIV Entry, Washington DC, USA, December 2007.
- *Geno2Pheno[coreceptor]*, Invited talk, Pfizer, Cologne, Germany, November 2007.
- *Geno2Pheno[coreceptor]*, Invited talk, Arevir Meeting 2007, Bonn, Germany, April 2007.
- *Genotypic and Phenotypic Analysis of Coreceptor Usage in a Large Cohort of Therapy-experienced Patients in Germany*, Invited talk, 4th International Workshop on Targeting HIV Entry, Rio Grande, Puerto Rico, December 2008 (presented by Thomas Lengauer due to illness).
- *Ultra-deep Sequencing Sata From the Bioinformatics Point of View - Challenges and Requirements for the Analysis*, Invited talk, 1st European Meeting on 454 Sequencing of Virus Populations, Penzberg, Germany, October 2008 (presented by André Altmann due to ilness).

Nils Weinhold:

- *Inferring Molecular Function from Local Function Conservation in Sequence and Structure Space*, Invited talk, ISMB 3D-SIG, Vienna, Austria, July 2007.

29.12 Teaching Activities

Summer Semester 2007

Courses:

- The Elements of Statistical Learning II (Lecturer: Thomas Lengauer, Tutor: Laura Tolosi)
- Biological Networks: Databases and Analysis Methods (Lecturer: Mario Albrecht, Tutors: Andreas Schlicker, Hagen Blankenburg, Dorothea Emig, Fidel Ramírez)

Winter Semester 2007/2008

Courses:

- The Elements of Statistical Learning I (Lecturer: Thomas Lengauer, Tutor: Jasmina Bogojeska)

Summer Semester 2008

Courses:

- Bioinformatics II (Lecturer: Thomas Lengauer, Tutor: Sven-Eric Schelhorn)

- The Elements of Statistical Learning II (Lecturer: Thomas Lengauer, Tutor: Jasmina Bogojeska)
- Biological Networks: Databases and Analysis Methods (Lecturer: Mario Albrecht, Tutors: Dorothea Emig, Alejandro Pironti)

Winter Semester 2008/2009

Courses:

- Structural Bioinformatics (Lecturer: Ingolf Sommer, Tutor: Jasmina Bogojeska)
- Molecular Networks in Biology and Medicine (Lecturer: Mario Albrecht at the University of Padova, Italy)

Diploma Theses

- Hagen Blankenburg: *A Distributed Annotation System for Molecular Interactions*, April 2007.
- Dorothea Emig: *The Impact of Alternative Splicing on Protein Interaction Networks*, May 2007.
- Martin Kircher: *In Silico Analysis of microRNA Genes on Human Chromosome 14 and Mouse Chromosome 12*, May 2007.
- Sven-Eric Schelhorn: *Prediction of Domain and Motif Interactions from Protein Networks*, June 2007.
- Andrea Volkamer: *Automated Generation of Pharmacophore Type Constraints to Improve FlexX Docking*, August 2007.
- Carla Haid: *BSIFlex: Binding Site Identification Including Protein Flexibility*, September 2007.
- Lars Feuerbach: *Towards Comparative Epigenomics: A Pilot Study on DNA Methylation of CpG Islands on Human Chromosome 21*, September 2007.
- Oliver Müller: *Using Shape Retrieval Techniques for Identifying Similar Protein Binding Sites*, September 2007.
- Hendrik Weisser: *Estimating the Local Variance of Protein Structure Models*, September 2007.
- Sandra Fischer: *Structure-based Prediction of HIV-1 Drug Resistance*, March 2008.
- Peter Schüffler: *MethMarker - A Toolkit for Design, Optimization and Validation of DNA Methylation Biomarkers for Cancer Diagnosis and Therapy Optimization*, April 2008.
- Matthias Dietzen: *An Additive Tree-based Approach to the Prediction of the Strength of Acidity for Drug-like Molecules*, August 2008.
- Yannick Djoumbou: *Molecular Dynamics based Prediction of HIV-1 Drug Resistance*, December 2008.
- Frederik Gwinner: *Assessing the Conservation of Protein Function on Evolutionary Defined Patches*, February 2009.

29.13 Dissertations, Habilitations, Offers, Awards

29.13.1 Dissertations

- Andreas Steffen, *Computational Approaches in Supramolecular Chemistry with a Special Focus on Virtual Screening*, January 2008
- Tobias Sing: *Model-based Anti HIV Therapy*, April 2008
- Christoph Hartmann: *Modeling of Flexible Side Chains for Protein-Ligand-Docking*, July 2008
- Christoph Bock: *Computational Epigenetics: Bioinformatic Methods for Epigenome Prediction, DNA Methylation Mapping and Cancer Epigenetics*, October 2008
- Priti Talwar: *Metabolic Network Analysis Based on Metabolomics Data*, November 2008
- Oliver Sander: *Structure Descriptor for the Analysis of Protein Structure, Function and Evolution*, December 2008
- Jochen Maydt: *Analysis of Recombination in Molecular Sequence Data*, December 2008

29.13.2 Offers for Faculty Positions

Mario Albrecht:

- W3-Professorship for Bioinformatics, University of Kiel, Germany

Iris Antes:

- Full Professorship in Bioinformatics, University of Bergen, Norway
- W2-Professorship for Protein Modelling and Computer-Aided Drug Design, Technical University of Munich, Germany

29.13.3 Awards

- *Thomas Lengauer* has been elected member of the German Academy of Science and Engineering - acatech in October 2007
- *Christoph Bock* has been awarded an Otto-Hahn Medal in the year 2009 for his dissertation. He also has received a Feodor Lynen Fellowship by the Alexander von Humboldt Foundation to pursue postdoctoral research at Harvard University and MIT Broad Institute, 2009–2011
- *Dorothea Emig* has received a travel grant by the Boehringer Ingelheim Fonds (Foundation for Basic Research in Medicine) for a research visit at the University of California, Berkeley, USA, February–April 2009
- *Konstantin Halachev* has received an EU travel grant (Marie Curie – Genome Architecture in Relation to Disease) to attend 3rd MC-CARD Conference, *Higher Order Genome Architecture*, April 2009
- *Christoph Hartmann* has received a DAAD travel grant for a research visit at the University of California, San Francisco, USA, September–October 2007

- *Lars Feuerbach* has received an EU travel grant (Marie Curie – Genome Architecture in Relation to Disease) to attend:
  - 2nd MC-CARD Conference, *Interplay among Genetics, Epigenetics and Non-coding RNA's*, May 2008
  - 2nd MC-CARD Workshop, *Genome Bioinformatic Techniques*, September 2008
  - 3rd MC-CARD Conference, *Higher Order Genome Architecture*, April 2009
- *Yassen Assenov* has received an EU travel grant (Marie Curie – Genome Architecture in Relation to Disease) to attend 3rd MC-CARD Conference, *Higher Order Genome Architecture*, April 2009

29.14 Grants and Cooperations

There are a number of projects funded with third-party money that are described in the succeeding subsections. We are listing projects that were funded during some time within the report period. We mention the cases in which the project has commenced or terminated within the report period.

DFG Projects

Cytochromes (partially funded through CBI)

In this project we are combining experimental and computational methods for the design of selective inhibitors for the human mitochondrial cytochrome P450 enzymes CYP11B1 and CYP11B2, respectively. These enzymes catalyze the final steps in the biosynthesis of cortisol and aldosterone (see Section 29.8.2). Funding for this project terminated in June 2007.

Partners:

- Max Planck Institute for Informatics (Dr. Iris Antes)
- Saarland University, Pharmacology (Prof. Rolf W. Hartmann, Ursula Müller, Sarah Ulmschneider)

Excellence Cluster on Multimodal Computing and Interaction

This project which has been funded since November 2007 is a major grant by DFG that encompasses all departments of the Institute and other Computer Science groups at the location. For this reason the partners are not listed here. The Department contributes to the cluster in the life science area (Research Area 6). One task to be tackled is the handling, quality management and visualization of complex biological data such as biological network data or high-dimensional statistical data as they arise in many fields in computational biology. Another task is to contribute to a demonstrator of the cluster, namely, a knowledge engine for a biological and medical domain. Our work on computational epigenetics and molecular networks fits into this setting.

Sequence Analysis for Hepatitis C Virus

This project is part of a DFG-funded Clinical Research Group on analyzing the hepatitis C virus with experimental and bioinformatics methods. In this project, we analyze the correlations between the genotypic variants of HCV and their phenotypic resistance behavior (see Sections 29.5.2 and 29.7.4).

Partners:

- DFG Clinical Research Group on Hepatitis C, directed by Prof. Stefan Zeuzem, Johann-Wolfgang-Goethe University Hospital, Frankfurt
- Max Planck Institute for Informatics (Dr. Christoph Welsch, Prof. Thomas Lengauer)

Structure and Interaction Analysis for Hepatitis C Virus

This project is part of a DFG-funded Clinical Research Group on analyzing the hepatitis C virus with experimental and bioinformatics methods. In this project, we analyze the structure, function, and interactions of HCV proteins (see Sections 29.5.2 and 29.7.4).

Partners:

- DFG Clinical Research Group on Hepatitis C, directed by Prof. Stefan Zeuzem, Johann-Wolfgang-Goethe University Hospital, Frankfurt
- Max Planck Institute for Informatics (Dr. Mario Albrecht, Prof. Thomas Lengauer)

BMBF Projects

Analysis of Expression Data

This project is funded by the NGFN (Nationales Genomforschungsnetzwerk). The NGFN Microarray Data Analysis Resource (`http://compdiag.molgen.mpg.de/ngfn/`) aimed at improving the bioinformatics and statistics support for the design and analysis of gene expression data in the NGFN. Basic techniques were taught in regularly held courses on the analysis of gene expression data. In addition, methods for enhancing biological interpretation of expression data (see Section 29.9.1) were developed by the Max Planck Institute for Informatics group in collaboration with Prof. Jörg Rahnenführer. Funding for this project terminated in May 2008.

Partners:

- Members of the NGFN SMP (Systematic Methodological Platform) Bioinformatics, directed by Prof. Roland Eils, DKFZ Heidelberg
- Department of Statists, University of Dortmund, Germany (Prof. Dr. Jörg Rahnenführer)
- Max Planck Institute for Informatics (Prof. Thomas Lengauer, Adrian Alexa)

HIV Cell Entry

This project is funded within the BMBF MedSys (Medical Systems Biology) Program. It is aimed at analyzing the effects of viral, cellular and pharmaceutical determinants of the effectiveness of viral cell entry of HIV both experimentally and through computational

modeling. Funding of this project commenced in March 2009. Thomas Lengauer coordinates the project.

Partners:

- Department of Virology, University of Heidelberg, (Prof. Hans-Georg Kräusslich)
- Institute of Virology, University of Cologne (Dr. Rolf Kaiser)
- Max Planck Institute for Informatics (Prof. Thomas Lengauer)

NGFN Bioinformatics Services for Environmental Diseases

This project was funded within NGFN (National Genome Research Network). Our task here was to provide structural and functional annotations for proteins that are targeted experimentally within the Genome Network on Environmental Diseases, see also Section 29.5.3). Funding for this project ended in May 2008 and was replaced with new funding from NGFNplus, see below.

Partners:

- Members of the NGFN Genome Network on Environmental Diseases, coordinated by Prof. Stefan Schreiber, University of Kiel
- Max Planck Institute for Informatics (Dr. Mario Albrecht, Prof. Thomas Lengauer)

NGFNplus Bioinformatics Support for Environmental Diseases

This project is funded within NGFN (National Genome Research Network). Our task is to provide insight into the structure, function, and interaction of disease proteins that are in the research focus of our partners within the Genome Network on Environmental Diseases, see also Section 29.5.3). Funding for this project started in January 2009.

Partners:

- Members of the NGFNplus Genome Network on Environmental Diseases, coordinated by Prof. Stefan Schreiber, University of Kiel
- Max Planck Institute for Informatics (Dr. Mario Albrecht)

NGFNplus Oncogene

This project, which is coordinated by Roman Thomas at the Max Planck Institute for Neurological Research in Cologne, has been funded since mid 2008. Its goal is a thorough investigation of molecular changes in cancer in order to identify prognostic markers and analyze therapy options. The project gathers experts from different fields of bioinformatics research, in order to optimally design and carry out wet lab experiments, analyze high-dimensional data with statistical methods, implement efficiently these methods, give biological meaning to novel discoveries and make best use of them to treat cancer patients.

Partners:

- Max Planck Institute for Neurological Research, Cologne (Dr. Roman Thomas)
- Cologne University (Prof. Dr. Peter Nürnberg)

- Cologne University Clinic (Prof. Dr. med. Jürgen Wolf)
- Düsseldorf University (Prof. Dr. Reza Ahmadian)
- Max Planck Institute of Molecular Physiology, Dortmund (Prof. Dr. Alfred Wittinghofer, Dr. Daniel Raugh, Prof. Dr. Herbert Waldmann)
- Dortmund University, Department of Statistics (Prof. Dr. Jörg Rahnenfürer)
- Max-Planck-Institute for Informatics (Prof. Dr. Thomas Lengauer)

EU Projects

BioSapiens

BioSapiens is the EU Network of Excellence for the bioinformatics annotation of the human genome. The network comprises 29 partners in Europe and is directed by Prof. Janet Thornton at the EBI (European Bioinformatics Institute). In this project, we are especially engaged in three work packages: (1) inferring protein function from sequence and structure, (2) bioinformatics for infectious diseases, and (3) DAS infrastructure for molecular interactions. We have also contributed to other work packages (ENCODE and cancer proteins).

Partners:

- Members of the Biosapiens Network

CancerDIP

Understanding the molecular events that initiate and maintain epigenetic gene silencing could lead to new clinical strategies for cancer prevention and therapy. The CancerDIP project aims to exploit new methods for genome-wide DNA methylation mapping, in order to develop epigenetic biomarkers for early diagnosis, prognosis and therapy optimization. CancerDIP is a European research initiative funded by the 7th Framework Programme for Research and Technological Development of the European Commission (FP7). It combines the expertise of six partners in five countries.

Partners:

- Fundación Centro Nacional de Investigaciones Oncológicas Carlos III, Madrid, Spain (Dr. Manel Esteller)
- Free University of Brussels, Brussels, Belgium (Dr. François Fuks)
- Radboud University, Nijmegen, The Netherlands (Prof. Henk Stunnenberg)
- Seconda Universitá degli Studi di Napoli, Naples, Italy (Prof. Lucia Altucci)
- Max Planck Institute for Informatics, Saarbrücken, Germany (Prof. Thomas Lengauer, Dr. Christoph Bock)
- Diagenode S.A., Liège, Belgium (Dr. Juana Magdalena)

CHAIN

This European integrated project whose funding is expected to start in the first half of 2009 brings together the major players in HIV resistance analysis in Europe. The EuResist consor-

tium is a partner in that project and, together with other partners, addresses Workpackage 4 (Bioinformatical analysis). There are too many partners in this project to be listed separately.

EuResist

The EuResist project aimed at developing a European integrated system for clinical management of antiretroviral drug resistance. The system provides the clinicians with a prediction of response to antiretroviral treatment in HIV patients, thus helping the clinicians to choose the best drugs and drug combinations for any given HIV genetic variant. To this end a European integrated data set has been created, linking some of the largest existing resistance databases. In this project, we have been especially engaged in the workpackages 3 (prediction models) and 4 (comparison and combination of models). The project terminated in September 2008, but the consortium stays on as an EEIG (European Economic Interest Group).

Partners:

- Informa S.r.l. (Francesca Incardorna)
- University of Siena (Prof. Maurizio Zazzi)
- Karolinska Institute (Prof. Anders Sönnerborg)
- Institute of Virology, University of Cologne (Dr. Rolf Kaiser)
- IBM Israel – science and technology LTD (Dr. Shai Fine)
- Max Planck Institute for Informatics (Prof. Thomas Lengauer)
- KFKI Research Institute for Particle and Nuclear Physics, Hungarian Academy of Sciences (Prof. Fulop Baszo)
- Kingston University (Dr. Andrea Petroczi)

Other Projects

Max-Planck-Fraunhofer Cluster on Machine Learning

This project brings together machine learning groups in Max-Planck Society and Fraunhofer Society to work on methods and applications. We bring our expertise of adapting and applying machine learning tools to computational biology into the project.

Partners:

- Max Planck Institute for Molecular Genetics, Berlin (Prof. Martin Vingron)
- Max Planck Institute for Biological Cybernetics, Tübingen (Prof. Bernhard Schölkopf)
- Max Planck Institute for Informatics, Saarbrücken (Prof. Thomas Lengauer)
- Fraunhofer Institute for Computer Architecture and Software Technology, Berlin (Prof. Klaus-Robert Müller)
- Fraunhofer Institute for Technical and Economical Mathematics, Kaiserslautern (Prof. Dieter Prätzel-Wolters)
- Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin (Prof. Stefan Wrobel)

– Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin (Prof. Martin Hofmann-Apitius)

29.15 Publications

Books

[1] T. Lengauer, ed. *Bioinformatics - From Genomes to Therapies 1. The Building Blocks: Molecular Sequences and Structures*. Wiley-VCH, Weinheim, Germany, 2007.

[2] T. Lengauer, ed. *Bioinformatics - From Genomes to Therapies 2. Getting at the Inner Workings: Molecular Interactions*. Wiley-VCH, Weinheim, Germany, 2007.

[3] T. Lengauer, ed. *Bioinformatics - From Genomes to Therapies 3. The Holy Grail: Molecular Function*. Wiley-VCH, Weinheim, Germany, 2007.

Journal articles and book chapters

[1] B. Adams, A. McHardy, C. Lundegaard, and T. Lengauer. Viral bioinformatics. In D. Frishman and A. Valencia, eds., *Modern Genome Annotation - The BioSapiens Network*, ch. 8.1, p. 23. Springer, Wien, 2009.

[2] G. Ahlenstiel, K. Roomp, M. Däumer, T. Nattermann, M. Vogel, J. Rockstroh, H. Beerenwinkel, R. Kaiser, H.-D. Nischalke, T. Sauerbruch, T. Lengauer, and U. Spengler. Selective pressures of HLA genotypes and antiviral therapy on human immunodeficiency virus type 1 sequence mutation at a population level. *Clinical and Vaccine Immunology*, 14(10):1266–1273, 2007.

[3] A. Altmann, N. Beerenwinkel, T. Sing, I. Savenkov, M. Däumer, R. Kaiser, S.-Y. Rhee, W. J. Fessel, R. W. Shafer, and T. Lengauer. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antiviral Therapy*, 12(2):169–178, 2007.

[4] A. Altmann, M. Däumer, N. Beerenwinkel, Y. Peres, E. Schülter, J. Büch, S.-Y. Rhee, A. Sönnerborg, W. J. Fessel, R. W. Shafer, M. Zazzi, R. Kaiser, and T. Lengauer. Predicting the response to combination antiretroviral therapy: Retrospective validation of geno2pheno-THEO on a large clinical database. *The Journal of Infectious Diseases*, 199:999–1006, 2009.

[5] A. Altmann, M. Rosen-Zvi, M. Prosperi, E. Aharoni, H. Neuvirth, A. Sönnerborg, E. Schülter, J. Büch, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi, and T. Lengauer. Comparison of classifier fusion methods for predicting response to anti HIV-1 therapy. *PLoS ONE*, 3(10):e3470, 2008.

[6] A. Altmann, T. Sing, H. Vermeiren, B. Winters, E. Van Craenenbroeck, K. Van der Borght, S.-Y. Rhee, R. W. Shafer, E. Schülter, R. Kaiser, Y. Peres, A. Sönnerborg, W. J. Fessel, F. Incardona, M. Zazzi, L. Bacheler, H. Van Vlijmen, and T. Lengauer. Advantages of predicted phenotypes and statistical learning models in inferring virological response to antiretroviral therapy from HIV genotype. *Antiviral Therapy*, 14(2):273–283, 2009.

[7] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, 2008.

[8] H. Blankenburg, R. D. Finn, A. Prlić, A. M. Jenkinson, F. Ramírez, D. Emig, S.-E. Schelhorn, J. Büch, T. Lengauer, and M. Albrecht. DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10):1321–1328, 2009.

[9] C. Bock. IVF: stars may have to consider the risk of stolen parenthood. *Nature*, 454(7207):938, 2008.

[10] C. Bock, K. Halachev, J. Büch, and T. Lengauer. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. *Genome Biology*, 10:R14, 2009.

[11] C. Bock and T. Lengauer. Computational epigenetics: Bioinformatik für neue wege in der krebsforschung. In *Jahrbuch 2007*, pp. 271–277. Max-Planck-Gesellschaft, München, 2007.

[12] C. Bock and T. Lengauer. Computational epigenetics. *Bioinformatics*, 24(1):1–10, 2008.

[13] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. CpG island mapping by epigenome prediction. *PLoS Computational Biology*, 3(6):1055–1070, 2007.

[14] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Research*, 36(10):e55, 2008.

[15] J. Bogojeska, A. Alexa, A. Altmann, T. Lengauer, and J. Rahnenführer. Rtreemix: an R package for estimating evolutionary pathways and genetic progression scores. *Bioinformatics*, 24(20):2391–2392, 2008.

[16] J. Bogojeska, T. Lengauer, and J. Rahnenführer. Stability analysis of mixtures of mutagenetic trees. *BMC Bioinformatics*, 9(1):165–181, 2008.

[17] F. Ceccherini-Silberstein, V. Svicher, T. Sing, A. Artese, M. M. Santoro, F. Forbici, A. Bertoli, S. Alcaro, G. Palamara, A. d. Monforte, J. Balzarini, A. Antinori, T. Lengauer, and C. F. Perno. Characterization and structural analysis of novel mutations in human immunodeficiency virus type 1 reverse transcriptase involved in the regulation of restistance to nonnucleoside inhibitors. *Journal of Virology*, 81(20):11507–11519, 2007.

[18] F. S. Domingues and T. Lengauer. Inferring protein function from protein structure. In T. Lengauer, ed., *Bioinformatics - From Genomes to Therapies 3. The Holy Grail: Molecular Function*, ch. 33, pp. 1211–1252. Wiley-VCH, Weinheim, Germany, 2007.

[19] F. S. Domingues, J. Rahnenführer, and T. Lengauer. Conformational analysis of alternative protein structures. *Bioinformatics*, 23(23):3131–3138, 2007.

[20] D. Emig, M. S. Cline, K. Klein, A. Kunert, P. Mutzel, T. Lengauer, and M. Albrecht. Integrative visual analysis of the effects of alternative splicing on protein domain interaction networks. *Journal of Integrative Bioinformatics*, 5(2):101–115, 2008.

[21] D. Emig, M. S. Cline, T. Lengauer, and M. Albrecht. Integrating expression data with domain interaction networks. *Bioinformatics*, 24(21):2546–2548, 2008.

[22] A. Flockerzi, J. Maydt, O. Frank, A. Ruggieri, E. Maldener, W. Seifarth, P. Medstrand, T. Lengauer, A. Meyerhans, C. Leib-Mosch, E. Meese, and J. Mayer. Expression pattern analysis of transcribed HERV sequences is complicated by ex vitvo recombination. *Retrovirology*, 4(1):39, 2007.

[23] A. Franke, T. Balschun, T. H. Karlsen, J. Sventoraityte, S. Nikolaus, G. Mayr, F. S. Domingues, M. Albrecht, M. Nothnagel, D. Ellinghaus, C. Sina, C. M. Onnie, R. K. Weersma, P. C. F. Stokkers, C. Wijmenga, M. Gazouli, D. Strachan, W. L. McArdle, S. Vermeire, P. Rutgeers, P. Rosenstiel, M. Krawczak, M. H. Vatn, the IBSEN study group, C. G. Mathew, and S. Schreiber. Sequence variants in IL10, ARPC2, and multiple other loci contribute to ulcerative colitis. *Nature Genetics*, 40(11):1319–1323, 2008.

[24] A. Franke, J. Hampe, P. Rosenstiel, C. Becker, F. Wagner, R. Häsler, R. D. Little, K. Huse, A. Ruether, T. Balschun, M. Wittig, A. ElSharawy, G. Mayr, M. Albrecht, N. J. Prescott, C. M. Onnie, H. Fournier, T. Keith, U. Radelof, M. Platzer, C. G. Mathew, M. Stoll, M. Krawczak, P. Nürnberg, and S. Schreiber. Systematic association mapping identifies NELL1 as a novel IBD disease gene. *PLoS ONE*, 2(8):e691.1–13, 2007.

[25] D. Frishman, M. Albrecht, H. Blankenburg, P. Bork, E. D. Harrington, H. Hermjakob, L. J. Jensen, D. A. Juan, T. Lengauer, P. Pagel, V. Schachter, and A. Valencia. Protein-protein interactions: analysis and prediction. In D. Frishman and A. Valencia, eds., *Modern Genome Annotation: The Biosapiens Network*, pp. 353–410. Springer, Wien, Austria, 2009.

[26] J. Hampe, A. Franke, P. Rosenstiel, A. Till, M. Teuber, K. Huse, M. Albrecht, G. Mayr, F. M. De La Vega, J. Briggs, S. Günther, N. J. Prescott, C. M. Onnie, R. Häsler, B. Sipos, U. R. Fölsch, T. Lengauer, M. Platzer, C. G. Mathew, M. Krawczak, and S. Schreiber. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nature Genetics*, 39(2):207–211, 2007.

[27] C. Hartmann, I. Antes, and T. Lengauer. IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Science*, 16:1294–1307, 2007.

[28] C. Hartmann, I. Antes, and T. Lengauer. Docking and scoring with alternative side-chain conformations. *Proteins: Structure, Function, and Bioinformatics*, 74(3):712–726, 2009.

[29] W. P. Hofmann, B. Fernandez, E. Herrmann, C. Welsch, U. Mihm, B. Kronenberg, G. Feldmann, U. Spengler, S. Zeuzem, and C. Sarrazin. Somatic hypermutation and mRNA expression levels of the bcl-6 gene in patients with hepatitis c virus-associated lymphoproliferative diseases. *Journal of Viral Hepatitis*, 14(7):484–491, 2007.

[30] W. P. Hofmann, C. Welsch, Y. Takahashi, H. Miyajima, U. Mihm, C. Krick, S. Zeuzem, and C. Sarrazin. Identification and in silico characterization of a novel compound heterozygosty associated with hereditary aceruloplasminemia. *Scandinavian Journal of Gastroenterology*, 42(9):1088–1094, 2007.

[31] M. Hornstein, M. J. Hoffmann, A. Alexa, M. Yamanaka, M. Müller, V. Jung, J. Rahnenführer, and W. A. Schulz. Protein phosphatase and TRAIL receptor genes as new candidate tumor genes on chromosome 8p in prostate cancer. *Cancer Genomics & Proteomics*, 5(2):123–136, 2008.

[32] A. M. Jenkinson, M. Albrecht, E. Birney, H. Blankenburg, T. Down, R. D. Finn, H. Hermjakob, T. J. P. Hubbard, R. C. Jimenez, P. Jones, A. Kähäri, E. Kulesha, J. R. Macías, G. A. Reeves, and A. Prlić. Integrating biological data - the Distributed Annotation System. *BMC Bioinformatics*, 9(Suppl. 8):S3.1–7, 2008.

[33] A. Kämper, D. Rognan, and T. Lengauer. Lead identification by virtual screening. In T. Lengauer, ed., *Bioinformatics - From Genomes to Therapies 2. Getting at the Inner Workings: Molecular Interactions*, ch. 18, pp. 651–704. Wiley-VCH, Weinheim, Germany, 2007.

[34] J. Kamradt, V. Jung, K. Wahrheit, L. Tolosi, J. Rahnenführer, M. Schilling, R. Walker, S. Davis, M. Stöckle, P. Meltzer, and B. Wullich. Detection of novel amplicons in prostate cancer by comprehensive genomic profiling of prostate cancer cell lines using oligonucleotide-based arrayCGH. *PLoS ONE*, 2(8):e769, 2007.

[35] R. Ketter, Y.-J. Kim, S. Storck, J. Rahnenführer, B. F. Romeike, W.-I. Steudel, K. D. Zang, and W. Henn. Hyperdiploidy defines a distinct cytogenetic entity of meningiomas. *Journal of Neuro-Oncology*, 83(2):213–221, 2007.

[36] R. Ketter, S. Urbschat, W. Henn, W. Feiden, N. Beerenwinkel, T. Lengauer, W.-I. Steudel, K. D. Zang, and J. Rahnenführer. Application of oncogenetic trees mixtures as a biostatistical model of the clonal cytogenetic evolution of meningiomas. *International Journal of Cancer*, 121(7):1473–1480, 2007.

[37] M. Kircher, C. Bock, and M. Paulsen. Structural conservation versus functional divergence of maternally expressed microRNAs in the Dlk1/Gtl2 imprinting region. *BMC Genomics*, 9:346, 2008.

[38] B. Kupfer, T. Sing, P. Schüffler, R. Hall, R. Kurz, A. McKeown, K.-E. Schneweis, W. Eberl, J. Oldenburg, H. H. Brackmann, J. K. Rockstroh, U. Spengler, M. Däumer, R. Kaiser, T. Lengauer, and B. Matz. Fifteen years of env C2V3C3 evolution in six individuals clonally infected with human immunodeficiency virus type 1. *Journal of Medical Virology*, 79(11):1629–1639, 2007.

[39] T. Lengauer. Bioinformatik. In H. Bullinger, ed., *Technologieführer - Grundlagen, Anwendungen, Trends*, p. 4. VID, Düsseldorf, 2007.

[40] T. Lengauer. Future trends. In T. Lengauer, ed., *Bioinformatics - From Genomes to Therapies 3. The Holy Grail: Molecular Function*, ch. 45, pp. 1651–1687. Wiley-VCH, Weinheim, Germany, 2007.

[41] T. Lengauer. Introduction. In T. Lengauer, ed., *Bioinformatics - From Genomes to Therapies 1. The Building Blocks: Molecular Sequences and Structures*, ch. 1, pp. 1–24. Wiley-VCH, Weinheim, Germany, 2007.

[42] T. Lengauer. Strategiekonzept "Molekulare Bioinformatik". In B. Reuse and R. Vollmar, eds., *Informatikforschung in Deutschland*, p. 9. Springer, Heidelberg, 2007.

[43] T. Lengauer, O. Sander, S. Sierra, A. Thielen, and R. Kaiser. Bioinformatics prediction of HIV coreceptor usage. *Nature Biotechnology*, 25(12):1407–1410, 2007.

[44] M. Lisurek, B. Simgen, I. Antes, and R. Bernhardt. Theoretical and experimental evaluation of a CYP106A2 low homology model and production of mutants with changed activity and selectivity of hydroxylation. *ChemBioChem*, 9(9):1439–1449, 2008.

[45] F. Liu, E. Tostesen, J. K. Sundet, T.-K. Jenssen, C. Bock, G. I. Jerstad, W. G. Thilly, and E. Hovig. The human genomic melting map. *PLoS Computational Biology*, 3(5):e93, 2007.

[46] A. J. Low, W. Dong, D. Chan, T. Sing, R. Swanstrom, M. Jensen, S. Pillai, B. Good, and P. R. Harrigan. Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. *AIDS*, 21(14):F19–F26, 2007.

[47] S. Lucas, R. Heim, M. Negri, I. Antes, C. Ries, K. E. Schewe, A. Bisi, S. Gobbi, and R. W. Hartmann. Novel aldosterone synthase inhibitors with extended carbocyclic skeleton by a combined ligand-based and structure-based drug design approach. *Journal of Medical Chemistry*, 51(19):6138–6149, 2008.

[48] G. Mayr, F. S. Domingues, and P. Lackner. Comparative analysis of protein structure alignments. *BMC Structural Biology*, 7:1, 2007.

[49] U. Mihm, O. Ackermann, C. Welsch, E. Herrmann, W.-P. Hofmann, N. Grigorian, H. Welker, T. Lengauer, S. Zeuzem, and C. Sarrazin. Clinical relevance of the 2'-5'-oligoadenylate synthetase/RNase L system for treatment response in chronic hepatitis c. *Journal of Hepatology*, 50(1):9, 2009.

[50] T. Mikeska, C. Bock, O. El-Maarri, A. Hübner, D. Ehrentraut, J. Schramm, J. Felsberg, P. Kahl, R. Büttner, T. Pietsch, and A. Waha. Optimization of quantitative MGMT promoter methylation analysis using pyrosequencing and combined bisulfite restriction analysis. *Journal of Molecular Diagnostics*, 9(3):368–381, 2007.

[51] D. Moser, S. Ekawardhani, R. Kumsta, H. Palmason, C. Bock, Z. Athanassiadou, K.-P. Lesch, and J. Meyer. Functional analysis of a potassium-chloride co-transporter 3 (SLC12A6) promoter polymorphism leading to an additional DNA methylation site. *Neuropsychopharmacology*, 34(2):458–467, 2009.

[52] J. Rahnenführer and T. Lengauer. Analysis of expression data: Classification of genes. In T. Lengauer, ed., *Bioinformatics - From Genomes to Therapies 2. Getting at the Inner Workings: Molecular Interactions*, ch. 27, pp. 993–1021. Wiley-VCH, Weinheim, Germany, 2007.

[53] F. Ramírez, A. Schlicker, Y. Assenov, T. Lengauer, and M. Albrecht. Computational analysis of human protein interaction networks. *Proteomics*, 7(15):2541–2552, 2007.

[54] K. Roomp. Evolution of drug resistance in HIV. In T. Lengauer, ed., *Bioinformatics - From Genomes to Therapies 3. The Holy Grail: Molecular Function*, ch. 40, pp. 1457–1496. Wiley-VCH, Weinheim, Germany, 2007.

[55] N. Salamat-Miller, J. Fang, C. W. Seidel, Y. Assenov, M. Albrecht, and C. R. Middaugh. A network-based analysis of polyanion-binding proteins utilizing human protein arrays. *Journal of Biological Chemistry*, 282(14):10153–10163, 2007.

[56] O. Sander, T. Sing, I. Sommer, A. J. Low, P. K. Cheung, P. R. Harrigan, T. Lengauer, and F. S. Domingues. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLoS Computational Biology*, 3(3):555–564, 2007.

[57] S.-E. Schelhorn, T. Lengauer, and M. Albrecht. An integrative approach for predicting interactions of protein regions. *Bioinformatics*, 24(16):i35–i41, 2008.

[58] N. Scheller, P. Resa-Infante, S. de la Luna, R. P. Galao, M. Albrecht, L. Kaestner, P. Lipp, T. Lengauer, A. Meyerhans, and D. Juana. Identification of PatL1, a human homolog to yeast P body component Pat1. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1773(12):1786–1792, 2007.

[59] A. Schlicker and M. Albrecht. FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Research*, 36(Database Issue):D434–D439, 2008.

[60] A. Schlicker, C. Huthmacher, F. Ramírez, T. Lengauer, and M. Albrecht. Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23(7):859–865, 2007.

[61] A. Schlicker, J. Rahnenführer, M. Albrecht, T. Lengauer, and F. S. Domingues. GOTax: investigating biological processes and biochemical activities along the taxonomic tree. *Genome Biology*, 8(3):R33.1–10, 2007.

[62] W. A. Schulz, A. Alexa, V. Jung, C. Hader, M. J. Hoffmann, M. Yamanaka, S. Fritzsche, A. Wlazlinski, M. Müller, T. Lengauer, R. Engers, A. R. Florl, B. Wullich, and J. Rahnenführer. Factor interaction analysis for chromosome 8 and DNA methylation alterations highlights innate immune response suppression and cytoskeletal changes in prostate cancer. *Molecular Cancer*, 6:1–16, 2007.

[63] S. Sierra, R. Kaiser, A. Thielen, and T. Lengauer. Genotypic coreceptor analysis. *European Journal of Medical Research*, 12(9):453–462, 2007.

[64] S. Sierra, R. Kaiser, A. Thielen, O. Sander, and T. Lengauer. Genotypic analysis of HIV co-receptor usage. In H. Jäger, ed., *Entry Inhibitoren - Neue Formen der HIV-Therapie*, ch. Rezeptorshift, pp. 31–39. Springer Medizin, Heidelberg, 2008.

[65] T. Sing, A. J. Low, N. Beerenwinkel, O. Sander, P. K. Cheung, F. S. Domingues, J. Büch, M. Däumer, R. Kaiser, T. Lengauer, and P. R. Harrigan. Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antiviral Therapy*, 12(7):1097–1106, 2007.

[66] K. Skrabal, A. J. Low, W. Dong, T. Sing, P. K. Cheung, F. Mammano, and P. R. Harrigan. Determining human immunodeficiency virus coreceptor use in a clinical setting: Degree of correlation between two phenotypic assays and a bioinformatic model. *Journal of Clinical Microbiology*, 45(2):279–284, 2007.

[67] C. Söderhäll, I. Marenholz, T. Kerscher, F. Rüschendorf, J. Esparza-Gordillo, M. Worm, C. Gruber, G. Mayr, M. Albrecht, K. Rohde, H. Schulz, U. Wahn, N. Hubner, and Y.-A. Lee. Variants in a novel epidermal collagen gene (COL29A1) are associated with atopic dermatitis. *PLoS Biology*, 5(9):e242.1952–1961, 2007.

[68] I. Sommer. Protein fold recognition based on distant homologs. In T. Lengauer, ed., *Bioinformatics - From Genomes to Therapies 1. The Building Blocks: Molecular Sequences and Structures*, ch. 11, pp. 351–388. Wiley-VCH, Weinheim, Germany, 2007.

[69] I. Sommer, O. Müller, F. S. Domingues, O. Sander, J. Weickert, and T. Lengauer. Moment invariants as shape recognition technique for comparing protein binding sites. *Bioinformatics*, 23(23):3139–3146, 2007.

[70] A. Steffen, M. Karasz, C. Thiele, T. Lengauer, A. Kämper, G. Wenz, and J. Apostolakis. Combined similarity and QSPR virtual screening for guest molecules of β-cyclodextrin. *New Journal of Chemistry*, 31(11):1941–1949, 2007.

[71] A. Steffen, C. Thiele, S. Tietze, C. Strassnig, A. Kämper, T. Lengauer, G. Wenz, and J. Apostolakis. Improved cyclodextrin-based receptors for camptothecin by inverse virtual screening. *Chemistry*, 13(24):6801–6809, 2007.

[72] M. L. Tress, P. L. Martelli, A. Frankish, G. A. Reeves, J. J. Wesselink, C. Yeats, P. Í. Ólason, M. Albrecht, H. Hegyi, A. Giorgetti, D. Raimondo, J. Lagarde, R. A. Laskowski, G. López, M. I. Sadowski, J. D. Watson, P. Fariselli, I. Rossi, A. Nagy, W. Kai, Z. Størling, M. Orsini, Y. Assenov, H. Blankenburg, C. Huthmacher, F. Ramírez, A. Schlicker, F. Denoued, P. Jones, S. Kerrien, S. Orchard, S. E. Antonarakis, A. Reymond, E. Birney, S. Brunak, R. Casadio, R. Guigo, J. Harrow, H. Hermjakob, D. T. Jones, T. Lengauer, C. A. Orengo, L. Patthy, J. M. Thornton, A. Tramontano, and A. Valencia. The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences*, 104(13):5495–5500, 2007.

[73] M. Trignetti, T. Sing, V. Svicher, M. Santoro, R. Forbici, R. D'arrigo, M. Bellocchi, M. Santoro, P. Marconi, M. Zaccarelli, M. Tortta, R. Bellagamba, P. Narciso, A. Antinori, C. Perno, T. Lengauer, and F. Ceccherini-Silberstein. Dynamics of NRTI resistance mutations during therapy interruption. *AIDS Research and Human Retroviruses*, 25(1):7, 2009.

[74] G. Van Ooijen, G. Mayr, M. Albrecht, B. J. C. Cornelissen, and F. L. W. Takken. Transcomplementation, but not physical association of the CC-NB-ARC and LRR domains of tomato R protein Mi-1.2 is altered by mutations in the ARC2 subdomain. *Molecular Plant*, 1(3):401–410, 2008.

[75] G. Van Ooijen, G. Mayr, M. M. A. Kasiem, M. Albrecht, B. J. C. Cornelissen, and F. L. W. Takken. Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. *Journal of Experimental Botany*, 59(6):1383–1397, 2008.

[76] J. D. Watson, J. M. Thornton, M. L. Tress, G. Lopez, A. Valencia, O. Redfern, C. A. Orengo, I. Sommer, and F. S. Domingues. Structure to function. In D. Frishman and A. Valencia, eds., *Modern Genome Annotation*, ch. 4.5, pp. 239–262. Springer, Vienna, 2008.

[77] N. Weinhold, O. Sander, F. S. Domingues, T. Lengauer, and I. Sommer. Local function conservation in sequence and structure space. *PLoS Computational Biology*, 4:e1000105, 2008.

[78] M. W. Welker, W.-P. Hofmann, C. Welsch, M. von Wagner, E. Herrmann, T. Lengauer, S. Zeuzem, and C. Sarrazin. Correlation nonstructural (ns)4b amino acid variations with initial viral kinetics during interferon-alpha-based therapy in hcv-1b-infected patients. *Journal of Viral Hepatitis*, 14(5):338–349, 2007.

[79] C. Welsch, M. Albrecht, J. Maydt, E. Herrmann, M. Welker, C. Sarrazin, A. Scheidig, T. Lengauer, and S. Zeuzem. Structural and functional comparison of the non-structural protein 4B of flaviviridae. *Journal of Molecular Graphics and Modeling*, 26(2):546–557, 2007.

[80] C. Welsch, F. S. Domingues, S. Susser, I. Antes, C. Hartmann, G. Mayr, A. Schlicker, C. Sarrazin, M. Albrecht, S. Zeuzem, and T. Lengauer. Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of HCV. *Genome Biology*, 9(1):R16, 2008.

[81] H. Zhu, I. Sommer, T. Lengauer, and F. S. Domingues. Alignment of non-covalent interactions at protein-protein interfaces. *PLoS ONE*, 3(4):e1926, 2008.

[82] E. Zotenko, J. Mestre, D. O'Leary, and T. Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLoS Computational Biology*, 4(8):e1000140, 2008.

Conference articles

[1] S. Bickel, J. Bogojeska, T. Lengauer, and T. Sheffer. Multi-task learning for HIV therapy screening. In A. McCallum and S. Roweis, eds., *Proceedings, Twenty-Fifth International Conference on Machine Learning*, Helsinki, Finland, 2008, pp. 56–63. Omnipress.

[2] D. Emig, K. Klein, A. Kunert, P. Mutzel, and M. Albrecht. Visualizing domain interaction networks and the impact of alternative splicing events. In *Proceedings of the 5th International Conference on BioMedical Visualization (MediVis): Information Visualization in Medical and Biomedical Informatics*, London, United Kingdom, 2008, pp. 36–43. IEEE Computer Society.

[3] L. Feuerbach, B. Christopher, K. Halachev, J. Büch, and T. Lengauer. Towards comparative epigenomics: A software toolkit for cross-species epigenome data analysis. In G. Meijer, T. Ried, W. Giaretti, and P. Hamilton, eds., *Abstracts of the 2nd MC-GARD Meeting*, Madrid, Spain, 2008, *CELLULAR ONCOLOGY Special Issue*, vol. 30, p. 230. IOS.

[4] M. Rosen-Zvi, A. Altmann, M. Prosperi, E. Aharoni, H. Neuvirth, A. Sönnerborg, E. Schülter, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi, and T. Lengauer. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. In *Proceedings of the 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2008)*, Toronto, Canada, 2008, *Bioinformatics*, vol. 24, pp. i399–406. Oxford University.

[5] H. Saigo and K. Tsuda. Iterative subgraph mining for principal component analysis. In F. Giannotti and D. Gunopulos, eds., *Proceedings of the IEEE International Conference on Data Mining (ICDM2008)*, Pisa, Italy, 2008, p. DM934. IEEE Computer Society.

[6] O. Sander, F. S. Domingues, H. Zhu, T. Lengauer, and I. Sommer. Structural descriptors of protein-protein binding sites. In A. Brazma, S. Miyano, and T. Akutsu, eds., *Proceedings of 6th Asia-Pacific Bioinformatics Conference*, Kyoto, Japan, 2008, pp. 79–88. Imperial College Press.

[7] T. Sing and N. Beerenwinkel. Mutagenetic tree Fisher kernel improves prediction of HIV drug resistance from viral genotype. In B. Schölkopf, J. Platt, and T. Hoffman, eds., *Advances in Neural Information Processing Systems 19*, Vancouver, B.C., Canada, 2007, pp. 1–9. MIT.

Theses

[1] I. Antes. *Computational Methods for the Investigation of Protein-Ligand Interactions*. Habilitation thesis, Universität des Saarlandes, 2009.

[2] H. Blankenburg. A distributed annotation system for molecular interactions. Diploma thesis, Ruprecht Karl Universität Heidelberg; Universität Heilbronn, 2007.

[3] C. Bock. *Computational Epigenetics - Bioinformatic methods for epigenome prediction, DNA methylation mapping and cancer epigenetics*. Phd thesis, Universität des Saarlandes, 2008.

[4] J. Bogojeska. Stability analysis of oncogenetic trees mixture models. Masters thesis, Universität des Saarlandes, 2007.

[5] M. Dietzen. An additive tree-based approach to the prediction of the strength of acidity for drug-like molecules. Masters thesis, Universität des Saarlandes, 2008.

[6] Y. Djoumbou. Molecular dynamics based prediction of hiv-1 drug resistance. Masters thesis, Universität des Saarlandes, 2008.

[7] D. Emig. The impact of alternative splicing on protein interaction networks. Diploma thesis, Eberhard Karls Universität Tübingen, 2007.

[8] L. Feuerbach. Towards comparative epigenomics: A pilot study on dna methylation of cpg islands on human chromosome 21. Masters thesis, Universität des Saarlandes, 2007.

[9] S. Fischer. Structure-based prediction of hiv-1 drug resistance. Masters thesis, Universität des Saarlandes, 2008.

[10] F. Gwinner. Assessing the conservation of protein function on evolutionary defined patches. Masters thesis, Universität des Saarlandes, 2009.

[11] C. Haid. Bsiflex: Binding site identification including protein flexibility. Masters thesis, Universität des Saarlandes, 2007.

[12] C. Hartmann. *Modeling of Flexible Side Chains for Protein-Ligand Docking*. Phd thesis, Universität des Saarlandes, 2008.

[13] M. Kircher. In silico analysis of microrna genes on human chromosome 14 and mouse chromosome 12. Masters thesis, Universität des Saarlandes, 2007.

[14] U. S. Korbel. Identification of symmetries in molecules and complexes. Masters thesis, Universität des Saarlandes, 2007.

[15] M. Mangels. In silico models for the prediction of metabolism by cytochrome p450 enzymes. Bachelor thesis, Universität des Saarlandes, 2009.

[16] J. Maydt. *Analysis of recombination in Molecular Sequence Data*. Phd thesis, Universität des Saarlandes, 2008.

[17] F. Müller. Inferring virological response to antiretroviral combination therapy based on past treatment lines. Bachelor thesis, Universität des Saarlandes, 2008.

[18] O. Müller. Using shape retrieval techniques for identifying similar protein binding sites. Masters thesis, Universität des Saarlandes, 2007.

[19] J. Perner. Semi-supervised learning for predicting anti-hiv drug resistance. Bachelor thesis, Universität des Saarlandes, 2008.

[20] O. Sander. *Structural Descriptors for the Analysis of Protein Structure, Function, and Evolution*. Phd thesis, Universität des Saarlandes, 2008.

[21] S.-E. Schelhorn. Prediction of domain and motif interactions from protein networks. Masters thesis, Universität des Saarlandes, 2007.

[22] P. Schüffler. Methmarker ? a toolkit for design, optimization and validation of dna methylation biomarkers for cancer diagnosis and therapy optimization. Masters thesis, Universität des Saarlandes, 2008.

[23] T. Sing. *Model-Based Anti HIV Therapy*. Phd thesis, Universität des Saarlandes, 2008.

[24] A. Steffen. *Computational Approaches in Supramolecular Chemistry with a special Focus on Virtual Screening*. Phd thesis, Universität des Saarlandes, 2008.

[25] D. Stöckel. Protein surface remeshing and characterisation using morse-smale complexes. Bachelor thesis, Universität des Saarlandes, 2008.

[26] P. Talwar. *Development of Computational Methods for Metabolic Network Analysis based on Metabolomics Data*. Phd thesis, Universität des Saarlandes, 2008.

[27] A. Volkamer. Automated generation of pharmacophore type constraints to improve flexx docking. Masters thesis, Universität des Saarlandes, 2007.

[28] H. Weisser. Estimating the local variance of protein structure models. Masters thesis, Universität des Saarlandes, 2007.

30 The Computer Graphics Group (D4)

30.1 Personnel

Director

Prof. Dr. Hans-Peter Seidel

Senior Researchers

Dr. Alexander Belyaev (until August 2007)
Dr. Elmar Eisemann (since October 2008)
Dr. Hendrik Lensch (until December 2008)
Dr. Meinard Müller (since September 2007)
Dr. Karol Myszkowski
Dr. Bodo Rosenhahn (until September 2008)
Dr. Robert Strzodka (since November 2007)
Dr. Thorsten Thormählen (since November 2007)
Dr. Michael Wand (since September 2007)

Researchers

Dr. Lionel Baboud (since March 2009)
Dr. Hongbo Fu (September 2008-January 2009)
Dr. Martin Fuchs
Dr. Thorsten Grosch (since September 2007)
Dr. Ruxandra Lasowski (since April 2008)
Dr. Sung Kil Lee (since February 2009)
Dr. Rafal Mantiuk (until April 2008)
Dr. Makoto Okabe (since April 2009)
Dr. Mike Sips (since September 2008)
Dr. Rhaleb Zayer (until April 2008)

PhD Students

Naveed Ahmed (until December 2008)
Lukas Ahrenberg (until September 2007)
Boris Ajdin (until April 2009)
Thomas Annen (until July 2008)
Tunc Aydin
Andreas Baak (since April 2008)
Robert Bargmann (until January 2008)
Tongbo Chen (until December 2008)

Edilson De Aguiar (until November 2008)
Piotr Didyk (since October 2008)
Zhao Dong
Christian Fuchs (until March 2009)
Jürgen Gall (until February 2009)
Miguel Andrés Granados Velásquez (since October 2008)
Peter Grosche (since June 2008)
Johannes Günther (until January 2008)
Nils Hasler
Thomas Helten (since October 2008)
Robert Herzog
Matthias Hullin (since April 2007)
Ivo Ihrke (until December 2007)
Jens Kerber (since October 2007)
Verena Konz (since May 2008)
Sergey Kosov (since September 2008)
Grzegorz Krawczyk (until October 2007)
Christian Kurz (since February 2008)
Torsten Langer (until April 2008)
Andrei Lintu (until November 2007)
Tobias Ritschel (since September 2007)
Waqar Saleem (until September 2008)
Natascha Sauber (until September 2007)
Oliver Schall (until July 2008)
Kristina Scherbaum
Volker Scholz (until July 2007)
Thomas Schultz
Mohammed Shaheen (since October 2008)
Kuangyu Shi (until March 2008)
Kaleigh Smith (until November 2008)
Carsten Stoll
Martin Sunkel
Art Tevs (since July 2007)
Wolfram von Funck (until September 2008)
Akiko Yoshida (until September 2008)
Gernot Ziegler (until June 2008)

Secretaries

Sabine Budde
Ellen Fries (since February 2009)
Cornelia Liegl (until December 2007)

30.2 Visitors

In the time period from April 2007 to April 2009, the following researchers visited our group:

Thorsten Grosch	21.05.07–22.05.07	Universität Koblenz, Koblenz, Germany
Marcus Magnor	21.05.07–22.05.07	TU Braunschweig, Braunschweig, Germany
Petar Dobrev	01.06.07–31.08.07	Universtiät Bremen, Bremen, Germany
Holger Theisel	14.06.07–15.06.07	Universtiät Bielefeld, Bielefeld, Germany
Kim Kil Joong	08.07.07–31.08.07	Seoul National University, Seoul, Korea
Holger Theisel	12.07.07–15.07.07	Universtiät Bielefeld, Bielefeld, Germany
Markus Magnor	15.07.07–16.07.07	TU Braunschweig, Braunschweig, Germany
Tamy Boubekeur	16.07.07–18.07.07	INRIA/University of Bordeaux, Bordeaux, France
Zachi Karni	31.07.07	HPL Israel, Haifa, Isreal
Georgia Albuquerque	09.09.07–10.09.07	TU Braunschweig, Braunschweig, Germany
Anna Tomaszewska	01.10.07–06.10.07	Institute of Technologies of Szczecin, Szczecin, Poland
Anna Tomaszewska	01.11.07–31.01.08	Institute of Technologies of Szczecin, Szczecin, Poland
Matthias Ihrke	05.10.07	Georg August Universität Göttingen, Göttingen, Germany
Kei Iwasaki	08.10.07–09.10.07	Wakayama University, Wakayama, Japan
Yoichi Sato	08.10.07–09.10.07	University of Tokyo, Tokyo, Japan
Tom Haber	15.10.07–31.01.08	Hasselt University, Diepenbeek, Belgium
Daniel Cremers	29.10.07	Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
Matthias Ihrke	01.11.07–31.01.07	Georg August Universität Göttingen, Göttingen, Germany
Reinhard Koch	06.11.07–23.11.07	Universität Kiel, Kiel, Germany
Michael Clausen	12.11.07–13.11.07	Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

Daniel Weinland	12.11.07–15.11.07	INRIA Grenoble Rhone-Alpes, Grenoble, France
Hartmut Schirmacher	15.11.07–17.11.07	Mercury Computer Systems GmbH, Berlin, Germany
Michael Gösele	21.11.07–22.11.07	TU Darmstadt, Darmstadt, Germany
Dorit Merhof	22.11.08	Universität Erlangen-Nürnberg, Erlangen, Germany
Andreas Baak	04.12.07–06.12.07	Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
Björn Krüger	04.12.07–06.12.07	Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
Elmar Eisemann	18.12.07	INRIA Rhône-Alpes, Saint Ismier Cedex, France
Michael Gösele	02.01.08–05.01.08	TU Darmstadt, Darmstadt, Germany
Michael Gösele	10.01.08–11.01.08	TU Darmstadt, Darmstadt, Germany
Christoph Garbe	15.01.08	Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany
Sebastian Ewert	20.01.08–22.01.08	Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
Michael Gösele	21.01.08–23.01.08	TU Darmstadt, Darmstadt, Germany
Carolin Früh	28.01.08–29.01.08	Universität Bern, Bern, Switzerland
Vlastimil Havran	09.03.08–16.03.08	Czech Technical University Prague, Prague, Czech Republic
Ruxandra Lasowski	11.03.08	Siemens Medical Solutions/TU München, München, Germany
Mikael Kalms	20.03.08	Linköping University, Linköping, Sweden
Hanno Ackermann	07.04.08–08.04.08	Okayama University, Okayama, Japan
Paul Debevec	15.04.08–17.04.08	University of Southern California, Los Angeles, USA
Hanno Ackermann	01.05.08–31.10.08	Okayama University, Okayama, Japan

Cristina Nader Vasconcelos	05.05.08–29.07.08	PUC Rio, Rio de Janeiro, Brasil
Oliver Grau	06.05.08	BBC Research & Innovation, Kingston upon Thames, UK
Tom Haber	01.06.08–31.07.08	Hasselt University, Diepenbeek, Belgium
Ramesh Raskar	06.06.08–07.06.08	MERL, Cambridge, MA, USA
Steve Marschner	18.06.08–22.06.08	Cornell University, Ithaca, NY, USA
Matthias Ihrke	30.06.08–03.07.08	Georg August Universität Göttingen, Göttingen, Germany
Jeppe Revall Frisvad	30.06.08–04.07.08	Technical University of Denmark, Kopenhagen, Dänemark
Richard Steffen	20.08.08–22.08.08	Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
Martin Bokeloh	01.09.08–30.09.08	Eberhard Karls Universität Tübingen, Tübingen, Germany
Alexander Berner	01.09.08–30.09.08	Eberhard Karls Universität Tübingen, Tübingen, Germany
Tom Haber	11.09.08–10.10.08	Hasselt University, Diepenbeek, Belgium
Mikael Kalms	15.09.08–14.12.08	Linköping University, Linköping, Sweden
Christian Linz	15.09.08–07.10.08	University of Southern California, Los Angeles, USA
Michael Gösele	22.09.08–24.09.08	TU Darmstadt, Darmstadt, Germany
Oliver Wang	01.11.08–15.03.09	University of California, Santa Cruz, USA
Sebastian Ewert	24.11.08–15.12.08	Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
Olaf Kähler	01.12.08–02.12.08	Friedrich-Schiller-Universität Jena, Jena, Germany
Martin Bokeloh	07.01.09–22.01.09	Eberhard Karls Universität Tübingen, Tübingen, Germany
Alexander Berner	07.01.09–22.01.09	Eberhard Karls Universität Tübingen, Tübingen, Germany
Michael Schwarz	17.02.09	Universität Erlangen-Nürnberg, Erlangen, Germany
Jeff Stuart	15.03.09–20.06.09	University of California, Davis, USA

Björn Schuller	17.03.09	TU München, München, Germany
Olivier Koch	26.03.09	MIT, Boston, MA, USA

30.3 Group Organization

Our research is currently organized into the following eight research areas, each having its own small group of coordinators:

- Digital Geometry Processing (M. Wand)
- Visualization (T. Schultz and M. Sips)
- Integrative Scientific Computing (R. Strzodka)
- Markerless Motion Capture and Multiview Stereo Processing (B. Rosenhahn and T. Thormählen)
- Multimedia Information Retrieval (M. Müller)
- General Appearance Acquisition and Computational Photography (H. Lensch)
- Advanced Global Illumination and Realtime Realistic Image Synthesis (E. Eisemann, T. Grosch and K. Myszkowski)
- High Dynamic Range Imaging and Perception Issues in Graphics (K. Myszkowski)

The coordinators coordinate the work in their areas together with Hans-Peter Seidel and they form the D4 steering committee. The steering committee meets on a weekly basis (Tuesday, 11 am) and discusses all group related issues. In particular, it addresses topics such as recruiting, guests and seminars, teaching, project acquisition, mid-term and long-term strategic planning.

The whole group meets thrice a week for the

- D4 lab meeting (Tuesday, 12:30 pm), where organizational issues are discussed and information is distributed by the members of the steering committee,
- D4 graphics colloquium (Tuesday, 1 pm), where visitors present their ongoing work to the group, the computer graphics group at Saarland University and to other interested people, and the
- D4 graphics lunch (Thursday, 11:30 am), where people from within D4 present their work in progress to the group. The main goal of this meeting is to keep the group informed on the ongoing projects, collect the group feedback and influence further project development at relatively early stages.

Apart from these formal meetings, there are several meetings and discussion groups that also take place frequently, but not on a totally regular basis, such as paper discussion groups that discuss papers of special interest, especially immediately preceding major conference events; technical meetings in special areas that are of particular interest to a specific subset of researchers (often in cooperation with people from the graphics group at Saarland University); internship and practical course meetings where all people involved in internships or FoPras meet and discuss; and last but not least meetings dedicated to single projects.

30.4 Digital Geometry Processing

Coordinator: Michael Wand

In the following, a number of novel algorithms for digital shape analysis and editing are presented. The first subsection (30.4.1) discusses new techniques for shape interrogation and deformation based on differential geometry tools. Afterwards, subsection 30.4.2 describes new techniques for the definition of smooth barycentric coordinates on surfaces. Subsection 30.4.3 proposes a new technique for noise reduction in 3D scanner data based on the idea of non-local filtering and Subsection 30.4.4 discusses mesh paramtrization. The following two sections deal with presentations of 3D geometry: View selection (subsection 30.4.5) and relief generation based on differential shape representations (subsection 30.4.6). Finally, the next three subsections (30.4.7-30.4.9) describe various techniques for solving geometric correspondence problems, in particular deformable shape matching and symmetry detection problems. Finally, we also have done work to interactively visualize and edit extremely large data sets such as large scale 3D scans; the results are presented in subsection 30.4.10.

30.4.1 Methods of Classical Differential Geometry for Shape Interrogation and Deformation

Investigators: Shin Yoshizawa, Torsten Langer, and Alexander Belyaev

Shape interrogation is the process of extraction of information from a geometric model. Shape interrogation methods are of increasing interest in geometric modeling and computer graphics. We give a general overview of shape interrogation approaches in [1].

A new geometry-based finite difference method for a fast and reliable detection of perceptually salient curvature extrema on surfaces approximated by dense triangle meshes is proposed in [4, 5]. The foundations of the method are two simple curvature and curvature derivative formulas overlooked in modern differential geometry textbooks and a new observation about inversion-invariant local surface-based differential forms.

In [3], a new free-form shape deformation approach is developed. We combine a skeleton-based mesh deformation technique with discrete differential coordinates in order to create natural-looking global shape deformations. Interesting links between the proposed free-form shape deformation technique and classical and modern results in the differential geometry of sphere congruences are established and discussed.

In [2] we use use the classical Euler's formula of the directional curvature to derive new formulas for the surface curvatures. The formulas are then used to obtain reliable estimates of the curvature tensor of a smooth surface approximated by a dense triangle mesh.

References

[1] S. Hahmann, A. Belyaev, L. Busé, G. Elber, B. Mourrain, and C. Rössl. *Shape Interrogation*, pp. 1–52. Springer, 2008.

- [2] T. Langer, A. Belyaev, and H.-P. Seidel. Exact and interpolatory quadratures for curvature tensor estimation. *Computer Aided Geometric Design*, 24(8-9):443–463, 2007.
- [3] S. Yoshizawa, A. Belyaev, and H.-P. Seidel. Skeleton-based variational mesh deformations. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 26(3):255–264, 2007.
- [4] S. Yoshizawa, A. Belyaev, H. Yokota, and H.-P. Seidel. Fast and faithful geometric algorithm for detecting crest lines on meshes. In M. Alexa, S. Gortler, and T. Ju, eds., *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, Maui, Hawaii, US, 2007, pp. 231–237. IEEE Computer Society.
- [5] S. Yoshizawa, A. Belyaev, H. Yokota, and H.-P. Seidel. Fast, robust, and faithful methods for detecting crest lines on meshes. *Computer Aided Geometric Design*, 25(8):545–560, 2008.

30.4.2 Generalized Barycentric Coordinates and their Applications

Investigators: Torsten Langer and Alexander Belyaev

One ever recurring topic in computer graphics is the need to express objects with respect to local coordinate systems. A special kind of such local coordinate systems are barycentric coordinate systems. While "standard" coordinate systems represent vectors as linear combination of certain basis vectors, barycentric coordinate systems represent points as linear combination of polytope vertices [1, 2].

Applications of barycentric coordinates range from shading over interpolation, finite elements applications, generalized Bézier surfaces, and parameterization methods to space deformations and dimensionality reduction. For some of these purposes the currently available "classical" barycentric coordinates are sufficient. However, many of the above mentioned applications are based on barycentric interpolation. For them, it is often desirable to have a *Hermite interpolation*. That is, we want to specify not only certain *values* at interpolation points but also the *derivatives*. For space deformations, this would allow to specify rotations and other linear transformations directly at a single point. So far, this had to be done by moving a whole group of control points.

We gave an axiomatic definition of *higher order barycentric coordinates* which can be used for Hermite interpolation [3]. Using them, we gain better local control over interpolations and less control points are necessary compared to previous methods. It turned out that this property is particular useful in the context of space deformations.

Our axiomatic definition of higher order barycentric coordinates defines not only a single set of these coordinates but a whole family of them. Of course, it is possible to construct such a coordinate system from scratch. However, the still ongoing research for good "classical" barycentric coordinates shows that this is a difficult research topic on its own. Therefore, we suggested to take a short-cut and construct higher order coordinates by modifying classical barycentric coordinates. We described an easy way for doing this. This allows to take advantage of existing implementations of these classical coordinates.

We demonstrate the extended interpolation capabilities of higher order coordinates using the example of higher order mean value coordinates. Mean value coordinates are defined

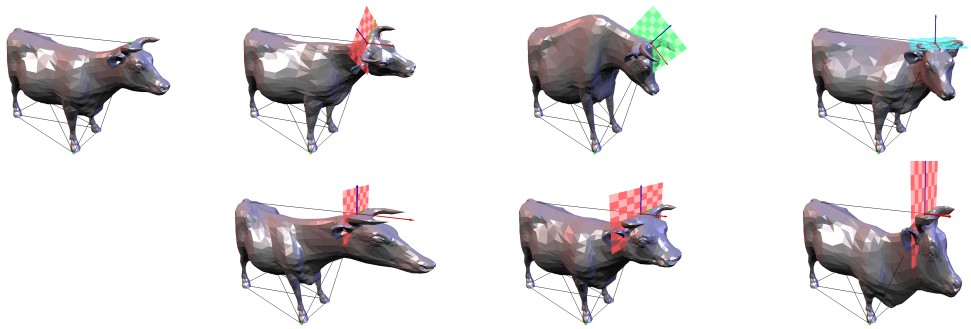


Figure 30.1: Changing single derivatives in a control net: rotations (top), non-uniform scaling (second row).

everywhere in $\mathbb{R}^3$, which turned out to be beneficial for our applications, and their main shortcomings vanish in the context of higher order coordinates while their benefits are retained.

In Figure 30.1, we did not perform any translations in order to focus on the deformations that are possible by altering the derivative at a single control point. Figure 30.2 shows how higher order mean value coordinates can be used for partial deformations. It is possible to construct a control net around the whole model where only those parts of it are "switched on" at a certain stage that are actually to be deformed while the remainder of the model remains unchanged. Figure 30.2 shows the basic building block for such a deformation system.

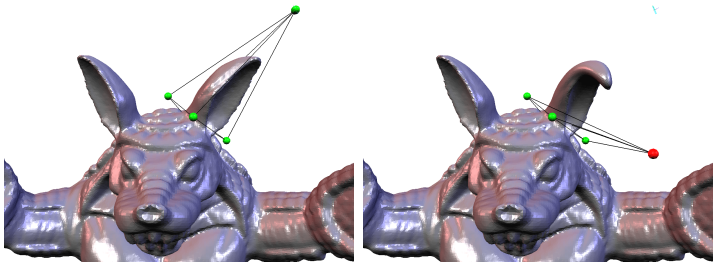


Figure 30.2: Partial shape deformation: Using derivative constraints with the higher order barycentric coordinates, we can ensure a smooth transition between deformed and undeformed parts of the shape, as shown here for the ear of the armadillo model.

We introduced a new type of barycentric coordinates which we called higher order barycentric coordinates because they allow to interpolate not only function values but linear functions. When used for space deformations, they introduce new means to manipulate objects. They can specify rotations and other linear transformations directly without the need to "simulate" such a transformation by moving a group of close-by control points. They can also be used to manipulate only parts of an object since the derivative constraints ensure a smooth transition between deformed and undeformed parts of the model.

Furthermore, we suggested a method to modify existing barycentric coordinates to create higher order barycentric coordinates. Therefore, they can be considered as a possible extension for existing coordinates rather than a completely new type. If we, nevertheless, compare higher order mean value coordinates and classical barycentric coordinates, higher order mean value coordinates are the only "shape-aware" non-negative, smooth barycentric coordinate functions, which are defined everywhere in $\mathbb{R}^n$, that we know of.

References

- [1] T. Langer, A. Belyaev, and H.-P. Seidel. Mean value coordinates for arbitrary spherical polygons and polyhedra in $\mathbb{R}^3$. In P. Chenin, T. Lyche, and L. L. Schumaker, eds., *Curve and Surface Design: Avignon 2006*, Avignon, France, 2007, Modern Methods in Mathematics, pp. 193–202. Nashboro.
- [2] T. Langer, A. Belyaev, and H.-P. Seidel. Mean value Bézier maps. In F. Chen and B. Jüttler, eds., *Advances in Geometric Modeling and Processing*, Hangzhou, China, 2008, *LNCS 4975*, pp. 231–243. Springer.
- [3] T. Langer and H.-P. Seidel. Higher order barycentric coordinates. In G. Drettakis and R. Scopigno, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Crete, Greece, 2008, vol. 27(2), pp. 459–466. Blackwell.

30.4.3 Geometry Denoising

Investigators: Oliver Schall, Rhaleb Zayer, and Alexander Belyaev

We have developed a new noise reduction technique for 3D scanner data based on the principle of non-local filtering [1, 2]. The filtering method focuses on noise removal on static and time-varying range data. The approach predicts the restored position of a perturbed vertex using similar vertices in its neighborhood. It defines the required similarity measure in a new *non-local* fashion which compares regions of the surface instead of point pairs. This allows the algorithm to obtain a more accurate denoising result than previous state-of-the-art approaches and, at the same time, to better preserve fine features of the surface. The method addresses the denoising problem differently compared to most previous approaches since it denoises range scans before they are combined within the scanning pipeline. This is more efficient since the given structure of the data can be utilized in a simpler similarity measure which allows for a faster evaluation. Furthermore, the approach extends more naturally to dynamic range data whose acquisition has become feasible thanks to interesting improvements in scanning technology. One example for the denoising efficiency of our approach is shown in Figure 30.3.

References

- [1] O. Schall, A. Belyaev, and H.-P. Seidel. Feature-preserving non-local denoising of static and time-varying range data. In B. Lévy and D. Manocha, eds., *ACM Symposium on Solid and Physical Modeling 2007*, Beijing, China, 2007, pp. 217–222. ACM.
- [2] O. Schall, A. G. Belyaev, and H.-P. Seidel. Adaptive feature-preserving non-local denoising of static and time-varying range data. *Computer-Aided Design*, 40(1):701–707, 2008.

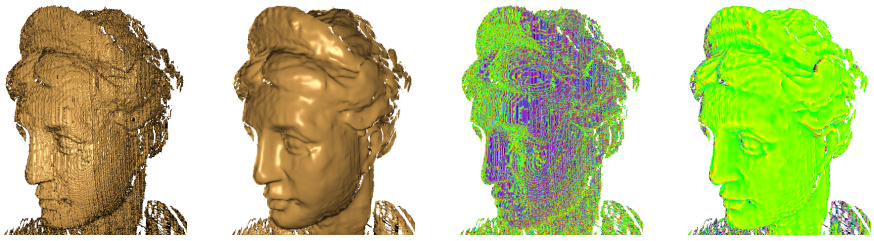


Figure 30.3: Denoising example: 3D scan of a bust.

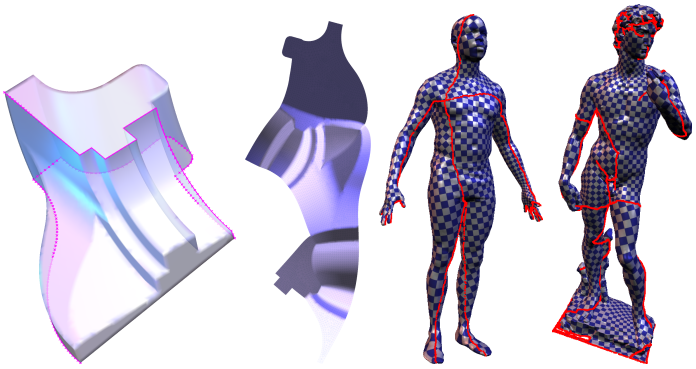


Figure 30.4: The original mechanical part model(left) and its planar parameterization obtained with the linear angle based parameterization. The texture mapping on the David and Man models reflects the quality of the parameterization.

30.4.4 Discrete Surface Parameterization and Quad Remeshing

Investigators: Rhaleb Zayer and Oliver Schall

Any attempt to flatten a non-trivial surface onto the plane will inevitably induce a certain amount of distortion. Surface parameterization is concerned with finding smooth one-to-one mapping between the surface and a parametric domain which yield minimal parametric distortion. One of the most prominent approaches in this direction is the Angle Based Flattening (ABF) which directly formulates the problem as a constrained nonlinear optimization in terms of angles. Since the original formulation of the ABF, a steady research effort has been dedicated to improving its efficiency. We reformulated the problem based on the notion of error of estimation. A careful manipulation of the resulting equations yields for the first time a linear version of angle based parameterization [2] which yields a speed up of two orders of magnitude in comparison to previous work. The error induced by this linearization is quadratic in terms of the error in angles and the validity of the approximation. Figure 30.4 shows some results obtained with our technique.

A related problem is remeshing, where we want to change the discretization of an object into primitives without altering its shape. A particularly challenging task is quad remeshing, where quadrilateral instead of triangular primitives have to be used. We have developed a new quad remeshing technique using vector fields defined over triangular meshes [1]. While the construction of such fields is by now a standard technique in geometry processing, enforcing

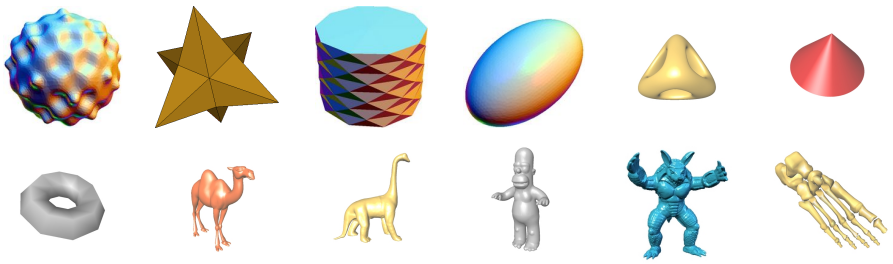


Figure 30.5: Shapes sorted from left to right and top to bottom according to one of our complexity measures. The least complex shape is on the top left and the most complex on the bottom right.

design constraints is still not fully investigated. We presented a technique which allows control over the gradient of a harmonic field by aligning it to a set of line constraints. The constraints can be sketched by the user or automatically obtained using a feature line detection algorithm. Furthermore, inspired by the problem of modeling heat flow on inhomogeneous surfaces, we investigated the potential of quasi-harmonic fields as a tool for controlling the behavior of the field over the surface. We demonstrated that it can be used for allowing certain regions on the surface to attract or repulse field contour lines. Both techniques can be used separately or together without affecting the computational cost since the Laplacian is a special case of the quasi-harmonic operator. In all cases the runtime is dominated by solving a single linear system.

References

- [1] O. Schall, R. Zayer, and H.-P. Seidel. Controlled field generation for quad-remeshing. In E. Haines and M. McGuire, eds., *ACM Symposium on Solid and Physical Modeling 2008*, Stony Brook, New York, USA, 2008, pp. 295–300. ACM.
- [2] R. Zayer, B. Lévy, and H.-P. Seidel. Linear angle based parameterization. In A. Belyaev and M. Garland, eds., *Symposium on Geometry Processing*, Barcelona, Spain, 2007, pp. 135–141. Eurographics/ACM.

30.4.5 Shape Annotating and View Selection: Static and Dynamic Views

Investigators: Waqar Saleem, Wenhao Song, and Alexander Belyaev

With rapid growth of Internet technology, digital libraries become a major source of information and data for scientists, educators, and students. One of the main activities of our group within AIM@SHAPE, a Network of Excellence project within EU's Sixth Framework Programme, consisted of maintaining and developing the project shape repository, http://shapes.aim-at-shape.net. This work greatly stimulated our research on shape retrieval, and representation. During the last two years, we studied the "best view" problem which recently attracted considerable attention from the pattern recognition and graphics communities – given a shape model, find view directions which deliver representative views of the shape.

We approach the above problem both in the static [3, 1] and dynamic [2] sense. In [3], we use a learning approach to automatically select a correct orientation for a shape and in [2] we extend the "best view" problem to compute a representative fly-by video instead of static images. We investigate in [1], methods and feasibility of transferring computed best view parameters from a shape to other similar shapes. We have also looked at the problem of determining the complexity of a shape from image information [4].

References

- [1] W. Saleem, G. Patanè, M. Spagnuolo, B. Falcidieno, and H.-P. Seidel. On continuously approximating discretely sampled view descriptors. Technical Report 7/2008, Istituto di matematica applicata e tecnologie informatiche (CNR-IMATI), Genova, Italy, 2008.
- [2] W. Saleem, W. Song, A. Belyaev, and H.-P. Seidel. On computing best fly. In M. Sbert, ed., *Proceedings of the 23rd Spring Conference on Computer Graphics, 2007*, Budmerice, Slovakia, 2007, pp. 143–149. Comenius University.
- [3] W. Saleem, D. Wang, A. Belyaev, and H.-P. Seidel. Automatic 2d shape orientation by example. In *2007 International Conference on Shape Modeling and Applications (SMI 2007)*, Lyon, France, 2007, pp. 221–225. IEEE.
- [4] D. Wang, A. Belyaev, W. Saleem, and H.-P. Seidel. Estimating complexity of 3d shapes using view similarity. Research Report MPI-I-2008-4-002, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, 2008.

30.4.6 Digital Bas-relief Generation

Investigators: Jens Kerber, Rhaleb Zayer, and Alexander Belyaev

Among all forms of sculpture, bas-relief is arguably the closest to painting. Although inherently a two dimensional sculpture, a bas-relief suggests a visual spatial extension of the scene in depth through the combination of composition, perspective, and shading.

The computation of a bas-relief, given a height field, can be regarded as a geometric counterpart of high-dynamic-range compression. We focus on compressing the depth interval size to the desired range in a way that visually important details are preserved. Therefore we proposed several different strategies which operate in the gradient domain. We extract the partial derivatives of the depth map, manipulate them by enhancing fine details and reassemble the final bas-relief from the new gradient field. Possible applications are of artistic nature like virtual sculpting, embossment, engraving or carving.

We developed three techniques which extend the works in this interesting young research area [1, 2]. We also reduced the necessary user intervention to a minimum and so the methods are more intuitive and user-friendly than before. Recently, we introduced more artistic freedom by inventing an automatic technique which allows to seamlessly combine different models or perspectives into one bas-relief in order to achieve cubism-like results.

References

- [1] J. Kerber, A. Belyaev, and H.-P. Seidel. Feature preserving depth compression of range images. In M. Sbert, ed., *Proceedings of the 23rd Spring Conference on Computer Graphics*, Budmerice,

Figure 30.6: All models are compressed to less than 2% of their initial spatial extend. Note the remaining details and the negligible depth which is demonstrated by the tilted views in the right most image.

Slovakia, 2007, pp. 110–114. Comenius University, Slovakia. Winner 2nd best SCCG 2007 paper award.

- [2] J. Kerber, A. Tevs, A. Belyaev, R. Zayer, and H.-P. Seidel. Feature sensitive bas relief generation. In *Proceedings of the IEEE International Conference on Shape Modeling and Applications (SMI) 2009*, 2009. IEEE. accepted for publication.

30.4.7 Isometric Registration of Ambiguous and Partial Data

Investigators: Art Tevs and Michael Wand

We introduce a new shape matching algorithm for computing correspondences between 3D surfaces that have undergone (approximately) isometric deformations. The new approach makes two main contributions: First, the algorithm is, unlike previous work, robust to "topological noise" such as large holes or "false connections", which is both observed frequently in real-world scanner data. Second, our algorithm samples the space of feasible solutions such that uncertainty in matching can be detected explicitly. We employ a novel randomized feature matching algorithm in order to find robust subsets of geodesics to verify isometric consistency. The paper shows shape matching results for real world and synthetic data sets that could not be handled using previous deformable matching algorithms.

We make three main contributions:

- We propose a randomized, RANSAC-like matching strategy that, unlike previous techniques, simultaneously estimates the correspondences and the validating geodesics in an outlier-robust way.
- We employ a new tangent-space optimization algorithm to optimize the placement of feature points for maximum isometric matching, yielding better result for noisy feature positions.
- We sample the space of plausible matches, which gives us the ability to explicitly examine matching alternatives. We consider this an important building block for fully automatic matching of complex models from several pieces.

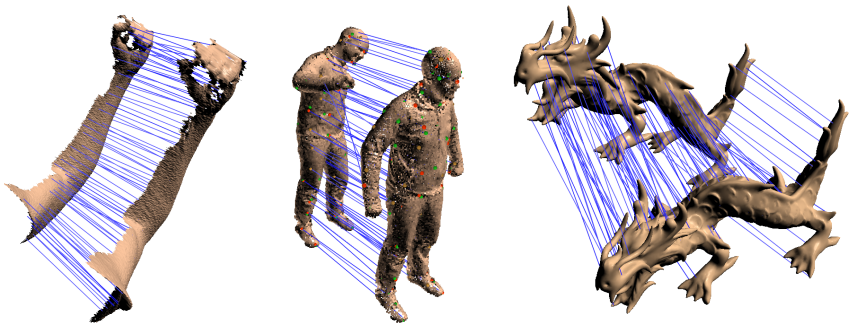


Figure 30.7: Different matching results of our technique [1].

The algorithm is, unlike previously known techniques, able to output matching alternatives by sampling the space of plausible solutions. This might be an important tool in multi part matching situations with ambiguous pairwise matches, such as animation sequence reconstruction. Even more this approach can be later used to perform symmetry detection in the point clouds, which we would like to investigate in the future work. Figures 30.7 - 30.8 show examples for correspondences obtained with our new matching technique.

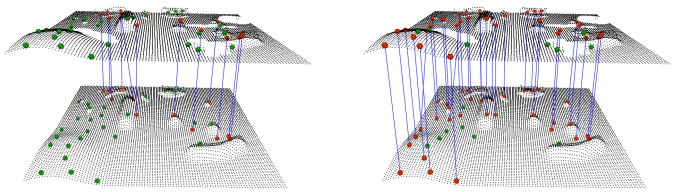


Figure 30.8: Result of matching of two different frames of a synthetic data set. Our technique [1] is robust to topological noise (i.e. acquisition holes). Left: Our technique without topological noise robustness. Right: Same technique with enabled robustness to topological noise.

References

- [1] A. Tevs, M. Bokeloh, M. Wand, A. Schilling, and H.-P. Seidel. Isometric registration of ambiguous and partial data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami Beach, Florida, USA, 2009. IEEE Computer Society.

30.4.8 Animation Reconstruction

Investigator: Michael Wand

Recently, a number of acquisition techniques have been proposed to record 3D geometry in real-time. This means that a scanning device acquires the movement and possibly deformation of a 3D object in real-time, like a 3D video camera. Such data is useful in various applications, such as special effects in movie productions or motion analysis for scientific purpose. A

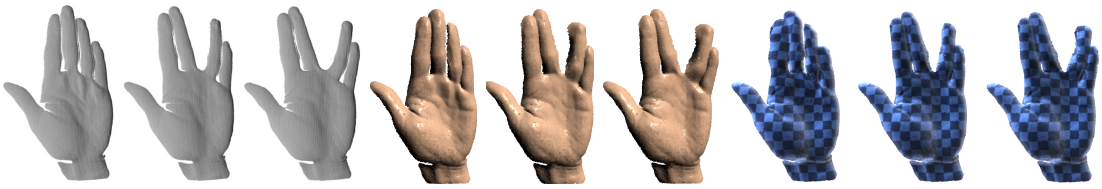


Figure 30.9: Example reconstruction results from a sequence of input scans of a deforming hand; **left:** original data points, **middle:** Reconstructed geometry, **right:** reconstructed correspondences indicated by a chessboard texture, with texture coordinates computed from correspondences.

real-time 3D scanner yields a number of unstructured clouds of measurement points, one for each acquisition time step. This data is typically noisy and not complete: It will show only the portion of the geometry that is visible to the scanner. In addition, the data does not provide any correspondence information over time; every measurement point is independent of all the others. This makes editing and postprocessing of such data difficult.

In order to overcome these problems, we have developed algorithms to automatically establish correspondences across animated point clouds and with this information fill in acquisition holes and reduce noise artifacts. Another way to formulate this problem is to factor the data into a common shape and its deformation. Our technique is based on a Bayesian reconstruction framework, extending earlier work on surface reconstruction ([2, 3]). The technique uses low level prior assumptions such as minimization of surface roughness, deformation and acceleration to obtain a variational model that allows fitting of a common, completed shape and a plausible deformation to the data points. The common shape itself is assembled incrementally from the data, thus not requiring an a priori template model [5]. In order to make this more efficient, we employ a subspace deformation model that discretizes the motion of the object on a coarser level than the geometry and a sweeping window solver, that can handle large sequences in an externally efficient way [4] (see Figure 30.9). A very similar computational framework based the same variational model also can be used to control animation sequences, thus allowing editing of animation sequences in a rather intuitive, example-based fashion [1].

References

- [1] B. Adams, M. Ovsjanikov, M. Wand, H.-P. Seidel, and L. Guibas. Meshless modeling of deformable shapes and their motion. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Dublin, Ireland, 2008, pp. 77–86. Eurographics Association.

- [2] Q.-X. Huang, B. Adams, and M. Wand. Bayesian surface reconstruction via iterative scan alignment to an optimized prototype. In D. Fellner and S. Spencer, eds., *SGP 2007, Fifth Eurographics Symposium on Geometry Processing*, Barcelona, Spain, 2007, pp. 213–223. Eurographics Association.

- [3] P. Jenke, M. Wand, and W. Straßer. Patch-graph reconstruction for piecewise smooth surfaces. In O. Deussen, D. Keim, and D. Saupe, eds., *Proceedings Vision, Modeling and Visualization (VMV 2008)*, Konstanz, Germany, 2008, pp. 3–12. Akademische Verlagsgesellschaft AKA.

- [4] M. Wand, B. Adams, M. Ovsjanikov, A. Berner, M. Bokeloh, P. Jenke, L. Guibas, H.-P. Seidel, and A. Schilling. Efficient reconstruction of non-rigid shape and motion from real-time 3d scanner data. *ACM Transactions on Graphics*, x, 2009. (to appear).

- [5] M. Wand, P. Jenke, Q.-X. Huang, M. Bokeloh, L. Guibas, and A. Schilling. Reconstruction of deforming geometry from time-varying point clouds. In A. G. Belyaev and M. Garland, eds., *Proc. 5th Eurographics Symposium on Geometry Processing (SGP 07)*, Barcelona, Spain, 2007, pp. 49–58. Eurographics Association.

30.4.9 Symmetry Detection

Investigators: Ruxandra Lasowski and Michael Wand

The previous two research topics are dealing with establishing correspondences between two different shapes. In addition, it is also interesting to examine correspondences within one and the same shape. This problem is usually named "symmetry detection" in literature and has recently gained a considerable amount of interest. The most commonly used techniques for symmetry detection are based on "transformation voting": First, a set of candidate correspondences (which might contain many outliers) is established. Second, mutual transformations are computed that align the candidate areas. By voting for such transformations, salient symmetries are found. This approach is elegant and efficient but has two major drawbacks: First, it cannot find a large number of simultaneous symmetries because this leads to a cluttering of transformation space that renders a stable extraction of symmetries impossible. Second, it is not easy to generalize this approach to more complex notions of symmetries (beyond simple transformations such as rigid mappings and scaling).

In this project, we have developed an alternative symmetry detection technique that is based on feature graphs: In a first step, we discretize the problem by detecting salient feature regions on the geometric object. We then connect these features with a spatial neighborhood graph, which is annotated with geometric quantities such as distances. In a second step, we examine this graph for similar subgraphs. For this step, we employ a randomized subgraph matching algorithm similar to the technique described in Section 30.4.7.

The graph-based technique indeed leads to a powerful and very general symmetry detection algorithm [1]. In particular, by designing good feature descriptors such as configurations of crease lines, we obtain a recognition performance ahead of previous state of the art techniques [2] (Figure 30.10).

References

- [1] A. Berner, M. Bokeloh, M. Wand, A. Schilling, and H.-P. Seidel. A graph-based approach to symmetry detection. In *Symposium on Volume and Point-Based Graphics*, Los Angeles, CA, 2008, pp. 1–8. Eurographics Association.

- [2] M. Bokeloh, A. Berner, M. Wand, H.-P. Seidel, and A. Schilling. Symmetry detection using line features. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28(2):697–706, 2009.

Figure 30.10: Two examples for symmetry detection; all symmetric instances are shown in the same color with dotted black and white lines to separate adjacent instances. Our algorithm is able to find a large number of symmetries within one piece of geometry. The data sets have been provided by the institute for cartography at Leibnitz University Hannover.

30.4.10 Editing Large Data Sets

Investigator: Michael Wand

We have also developed tools to handle very large geometric data sets. This is an important prerequisite to work with the large amounts of data that for example originates from 3D scanning devices. In particular, we have developed a fully dynamic multi-resolution data structure for point sampled geometry that allows for the first time both a real-time visualization and real-time editing of large 3D scanner data sets, with complexity only bound by available hard drive space [2, 1, 3]. We have also developed a software system that virtualizes the access to large geometric data bases using this approach and is used as a basis for many of the other research projects described in this section. A more detailed description of the software framework is given later in section 30.12.4.

References

- [1] M. Wand. Rendering of very large models. In M. Gross and H. Pfister, eds., *Point-Based Graphics*, pp. 313–326. Morgan Kaufmann/Elsevier, Amsterdam, The Netherlands, 2007.
- [2] M. Wand, A. Berner, M. Bokeloh, A. Fleck, M. Hoffmann, P. Jenke, B. Maier, D. Staneker, and A. Schilling. Interactive editing of large point clouds. In B. Chen, M. Zwicker, M. Botsch, and R. Pajarola, eds., *Symposium on Point-Based Graphics 2007, Eurographics / IEEE VGTC Symposium Proceedings*, Prague, Czech Republik, 2007, pp. 37–46. Eurographics Association.
- [3] M. Wand, A. Berner, M. Bokeloh, P. Jenke, A. Fleck, M. Hoffmann, B. Maier, D. Staneker, A. Schilling, and H.-P. Seidel. Processing and interactive editing of huge point clouds from 3d scanners. *Computers and Graphics*, 32(2):204–220, 2008.

30.5 Visualization

Coordinators: Thomas Schultz and Mike Sips

The goal of visualization is to produce visual representations that facilitate the exploration, understanding, and communication of large and complex datasets. We have performed research

both in scientific visualization, which concentrates on measurement and simulation data on spatial grids, and information visualization, which deals with data with a large number of abstract dimensions.

In scientific visualization, our focus was on medical data from diffusion weighted magnetic resonance imaging, on data from computational fluid dynamics, and on the application of visualization techniques to shape deformations. In information visualization, our focus was to study principles of effective visualizations of high-dimensional data spaces and geo-spatial data. A better understanding of these principles are crucial for the automated generation of graphical representations of databases.

30.5.1 Feature Extraction for DW-MRI Visualization

Investigators: Thomas Schultz, Natascha Sauber, and Holger Theisel

Diffusion Weighted Magnetic Resonance Imaging (DW-MRI) is a medical imaging modality that offers the unique opportunity to investigate the white matter of the human brain non-invasively. However, it produces large amounts of volumetric data which are impossible to analyze for human experts without adequate preprocessing and visualization. Our general strategy to address this complexity is to extract and to visualize meaningful features.

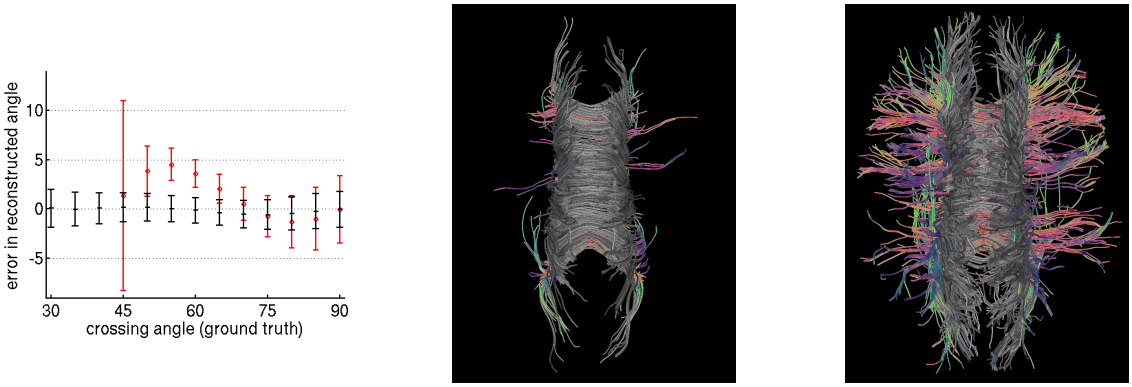
In our work [3], we demonstrate that an existing generalization of vector field topology to generic symmetric tensor fields does not produce useful results when applied to diffusion tensor (DT-MRI) data. Instead, novel topological features are proposed which are founded on the anatomical meaning of the data and reflect the uncertainty inherent in any connectivity estimate from diffusion imaging.

Often, diffusion images are not the only type of data acquired from a subject. In cooperation with the Max Planck Institute for Human Cognitive and Brain Sciences (Leipzig), we developed a method to visually integrate streamlines from DT-MRI fiber tracking with context from structural MRI [1]. The method is modeled on an anatomical fiber preparation technique known as Klingler dissection. It also demonstrates that opacity isosurfaces are expressive features in scalar fields and that they can be extracted and rendered efficiently.

Many types of features in DW-MRI data require to estimate the number, orientations, and volume fractions of individual nerve fiber tracts within a voxel. Our contribution in [2] is to improve the reliability of such estimates in cases where high angular resolution data is used to investigate crossing fiber bundles. This is achieved by a low-rank approximation of higher order tensors, which are used to model the orientation distribution functions which stem from Q-Ball imaging and spherical deconvolution. Figure 30.11 presents examples of the improved results.

References

- [1] T. Schultz, N. Sauber, A. Anwander, H. Theisel, and H.-P. Seidel. Virtual klingler dissection: Putting fibers into context. *Computer Graphics Forum (Proc. EuroVis)*, 27(3):1063–1070, 2008.
- [2] T. Schultz and H.-P. Seidel. Estimating crossing fibers: A tensor decomposition approach. *IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE Visualization)*, 14(6):1635–1642, 2008.



(a) (b) (c)

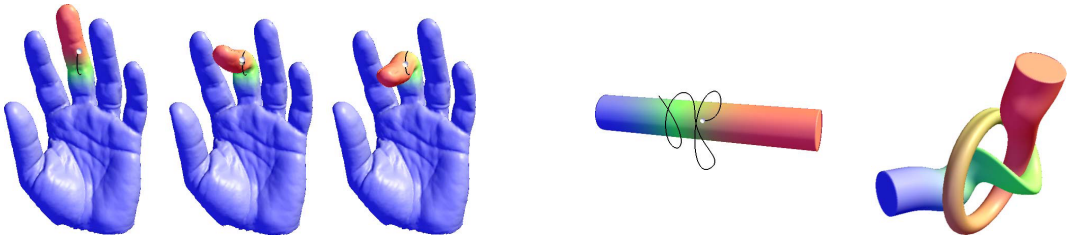
Figure 30.11: In synthetic data (a), tensor approximation (black) reconstructs crossing nerve fibers over a wider angular range and with less bias than traditional maximum search (red). In real data, it allows for reconstruction of the transcallosal fibers (red in c), which are not found with previous methods (b).

- [3] T. Schultz, H. Theisel, and H.-P. Seidel. Topological visualization of brain diffusion MRI data. *IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE Visualization)*, 13(6):1496–1503, 2007.

30.5.2 Vector Field Based Shape Deformations

Investigators: Wolfram von Funck and Holger Theisel

In the previous report, we presented an approach in which insights from vector field processing and visualization were applied to perform shape deformations. To this end, we construct C^1 continuous time-dependent vector fields and perform a path line integration of mesh vertices. The fact that the we construct the fields to be divergence-free guarantees that the deformation is volume-preserving and intersection-free. We now extended this basic approach to make it more usable and applicable to a wider range of problems.



(a) (b)

Figure 30.12: With explicit boundary control, we can deform a finger without affecting the rest of the hand (a). Specifying the deformation as a path allows for strong deformations like knots (b).

A drawback of the original method was the way the deformations were specified, using simple, but inflexible volumetric tools like spheres and cylinders. In [2], we allow the user to mark the regions of deformation directly on the surface instead (red area in Figure 30.12 (a)). The deformation itself is then specified as the path of a handle, along with a scalar function that defines a twisting. The full path of the handle is taken into account, which allows for extreme deformations like knots (cf. Figure 30.12 (b)).

An effect that helps to make animations appear more realistic are secondary deformations, small changes of the shape that result from the larger-scale primary deformation. An example of such deformations are jiggling and bouncing effects on the skin of a moving human. In [1], we demonstrate that plausible elastic secondary deformations can be derived via a simple mass-spring model, which provides parameters for constructing divergence-free vector fields. In contrast to the original approach, the resulting vector fields are C^2 rather than C^1, and can be described by polynomials of lower degree. Moreover, an improved GPU-based implementation allows us to treat even larger models interactively.

References

- [1] W. von Funck, H. Theisel, and H.-P. Seidel. Elastic secondary deformations by vector field integration. In A. Belyaev and M. Garland, eds., *Proceedings of the 5th Eurographics Symposium on Geometry Processing*, Barcelona, Spain, 2007, *ACM International Conference Proceeding Series*, vol. 257, pp. 99–108. ACM.
- [2] W. von Funck, H. Theisel, and H.-P. Seidel. Explicit control of vector field based shape deformations. In M. Alexa, S. Gortler, and T. Ju, eds., *Proceedings of the Pacific Conference on Computer Graphics and Applications, Pacific Graphics 2007*, Maui, Hawaii, 2007, pp. 291–300. IEEE Computer Society.

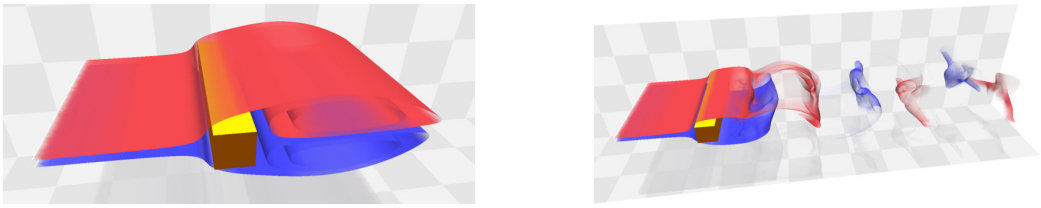
30.5.3 Interactive Vector Field Visualization

Investigators: Thomas Annen, Wolfram von Funck, and Holger Theisel

Three-dimensional vector fields are a common result of computational fluid dynamics. Due to their high information content, straightforward generalizations of established methods for the visualization of two-dimensional flow fields, like line integral convolution and extraction of topological skeletons, produce visual clutter. This complexity increases even further when the field is time- or parameter-dependent. Rather than trying to capture the full data in a single rendering, our strategy is to allow for an interactive exploration.

Contours have proven to be an effective tool to convey surfaces and three-dimensional scalar fields. Even though they do not have a straightforward generalization to vector-valued data, we demonstrate in [1] that the basic idea of using contours can be transferred to vector fields: Based on the current viewing direction, we define seeding and termination criteria for a streamline-based visualization. A GPU-based implementation ensures that the viewpoint can be changed at interactive frame rates, and our mathematical formulation guarantees temporal coherence.

Our implementation of vector field contours aims at static 3D vector fields. In experimental flow visualization, a common method to visualize time-dependent flow is to inject smoke into the volume, e.g., from a burning stick, or from a smoke nozzle. Inspired by the observation



(a) (b)

Figure 30.13: Smoke surfaces can be used to depict the dynamical flow behind a cube. Subfigure (a) shows the recirculation bubble at an early stage of the simulation. After vortex shedding sets in, the flow starts to exhibit true three-dimensional behavior (b).

that smoke has a surface-like appearance in artistic photographs, our approach [2] renders smoke as semi-transparent streak surfaces (Figure 30.13). Modulating opacity with triangle area and shape allows us to avoid expensive remeshing and leads to interactive surface computation and rendering. This makes it possible to explore time-dependent flows in an intuitive manner.

References

- [1] T. Annen, H. Theisel, C. Rössl, G. Ziegler, and H.-P. Seidel. Vector field contours. In C. Shaw and L. Bartram, eds., *Proceedings of the Graphics Interface 2008*, Windsor, Ontario, Canada, 2008, pp. 97–105. ACM.
- [2] W. von Funck, T. Weinkauf, H. Theisel, and H.-P. Seidel. Smoke surfaces: An interactive flow visualization technique inspired by real-world flow experiments. *IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE Visualization)*, 14(6):1396–1403, 2008.

30.5.4 Path Line Oriented Flow Topology on Time-dependent Flow Fields

Investigators: Kuangyu Shi and Holger Theisel

Topological methods reduce flow fields to their qualitative structure. In static flows, streamlines describe the trajectories of massless particles. Topological methods partition the domain into regions of uniform asymptotic flow behavior: In each face of a topological skeleton, streamlines start in the same source and end in the same sink. Time-dependent flows are traditionally addressed by considering the streamline topology of individual time steps, and tracking the resulting skeleton over time.

Our project concentrates on pathlines instead, since they describe the trajectories of massless particles in time-dependent flows. Pathlines are typically defined over a limited time interval, so their asymptotic behavior cannot be investigated. Instead, we consider a reformulation of flow topology, in which the source and sink of a streamline are considered streamline attributes. In this sense, we transfer topological methods to pathlines by computing a variety of pathline attributes, and visualizing them with state-of-the-art methods from information visualization [1]. This approach was used successfully to identify relevant structures in three different three-dimensional, time-dependent flow datasets.

The finite time Lyapunov exponent (FTLE) measures the spread of infinitesimally close particles after finite integration time. Lagrangian coherent structures (LCS) are ridges in the FTLE field, and are comparable to topological separatrices in that they indicate transport barriers in time-dependent flow. Many visualization methods either convey the dynamics of the flow (like pathline trajectories) or local quantitative properties (like kinetic energy or momentum). We propose to combine both aspects by filtering scalar quantities along pathlines, in a manner similar to line integral convolution (LIC) [2]. Interestingly, the resulting structures resembled FTLE fields in many cases.

References

- [1] K. Shi, H. Theisel, H. Hauser, T. Weinkauf, K. Matkovic, H.-C. Hege, and H.-P. Seidel. Path line attributes - an information visualization approach to analyzing the dynamic behavior of 3D time-dependent flow fields. In H.-C. Hege, K. Polthier, and G. Scheuermann, eds., *Topology-Based Methods in Visualization II*, Grimma, Germany, 2009, Springer series of Mathematics and Visualization, pp. 75–88. Springer.

- [2] K. Shi, H. Theisel, T. Weinkauf, H.-C. Hege, and H.-P. Seidel. Finite-time transport structures of flow fields. In *Proceedings of the IEEE VGTC Pacific Visualization Symposium (Pacific Vis 2008)*, Kyoto, Japan, 2008, pp. 63–70. IEEE.

30.5.5 Visual Analytics in High-dimensional Data Spaces

Investigator: Mike Sips

A major challenge is how to present high dimensional data to the analyst. Many visualization methods involve mapping high dimensional data to lower-dimensional views. Because graphical displays are composed of two spatial coordinates and a limited number of visual variables such as color, texture, etc., the maximum number of dimensions that can be shown in anyone view is roughly 3-8. And since the dimensionality of the data is often quite high – often tens to hundreds of dimensions – the mapping from data space to display space involves a loss of information.

Together with Pat Hanrahan and John Lewis at Stanford University, we introduced class consistency as a criterion for ranking and selecting good views to a class model from among the numerous possible projections of a high-dimensional data set [2]. Class consistency characterizes the extent to which the class neighborhood structure in the high-dimensional data is preserved in a low-dimensional view. This method can be applied to data with preexisting categorical labels, or to data that has been organized into classes with a clustering algorithm.

We proposed two computable measures of consistency. The first, distance consistency, is easy to implement and is well suited for data with convex clusters. We found that this measure is correlated with people's preferred views of a variety of real world data sets. The second measure, distribution consistency, is more general and can assess non-convex and interleaved data distributions. We compared these two measures on a variety of data sets and found that they were highly correlated. The use of these consistency measures can reduce or eliminate the need for the analyst to manually search among a large number of data projections (see Figure 30.14 for an illustration).

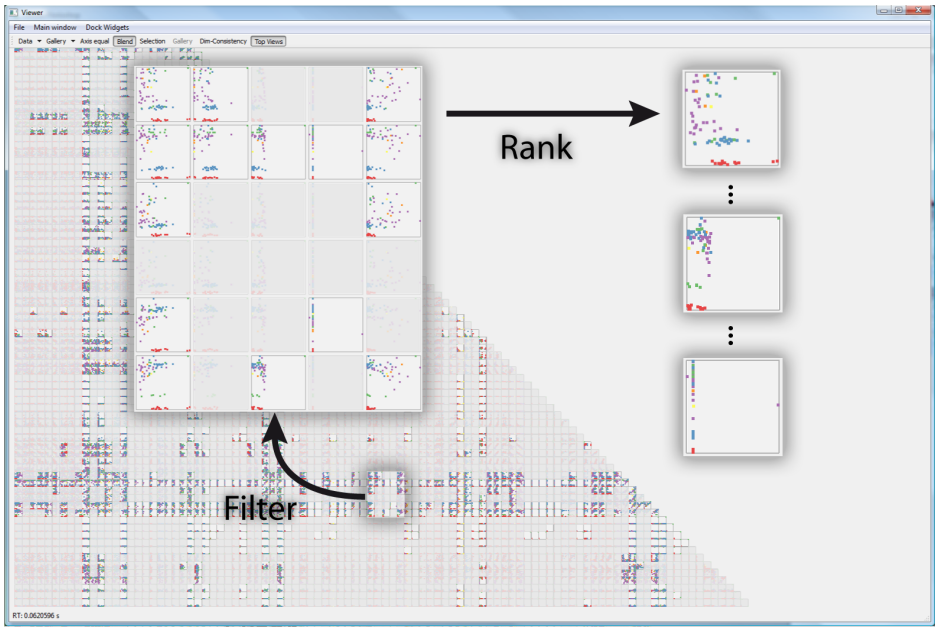


Figure 30.14: The interactive threshold sliders is used to fade out poor views in order to support the human analyst to find understand a class structure in n–D. The WHO data sets consists of 159 dimensions which results in over 12.000 unique 2–D scatterplots, and each data point has been labeled belonging to a HIV risk group. In this example, views below a consistency threshold of 80 are faded out. Many irrelevant views are faded-out and the number of views to look at can be interactively reduced to a manageable size. Note the upper triangular matrix of 2–D scatterplots is omitted.

In my research at the Max-Planck-Institute, we extended our filtering and ranking interface to collaborative analysis work scenarios. The problem here is that the hardware diversity of collaborative analysis settings in modern business intelligence or science centers – so-called smart rooms– typically ranges from handheld devices to wall-sized displays. The effectiveness of a visualization, however, significantly varies across different display devices. This may cause incorrect interpretations of the data because an analyst usually has no visual clues as to which extent a class structure is faithfully reproduced on different display devices.

We also proposed a novel method for determining meaningful parameter- and attribute settings for visualizations. During the last two decades, a wide variety of advanced methods for the visual exploration of large data sets have been proposed. For most of these techniques user interaction has become a crucial element, since there are many situations in which a user or an analyst has to select the right parameter settings from among many or selects a subset of the available attribute space for the visualization process. Our technique called Pixnostics [1] automatically analyzes pixel images that are the result from diverse parameter mappings and ranks them according to the potential value for the user. This allows a more effective and more efficient visual data analysis process, since the attribute/parameter space is reduced to meaningful selections, and thus the analyst obtains faster insight into the data. We demonstrated the benefit of our approach in real world applications.

References

- [1] J. Schneidewind, M. Sips, and D. Keim. An automated approach for the optimization of pixel-based visualizations. *Information Visualization*, 6(1):75–88, 2007.
- [2] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum (Proc. EuroVis 2009)*, 28(3), 2009.

30.5.6 Geo-Spatial Visualization

Investigator: Mike Sips

Value-by-area cartograms attracted the interest of many artists [6] [5] and cartographer [4] [1]. They emphasize the relative importance of geographical regions through re-sizing the display area of continents, nations, federal states, etc. proportionate to a given statistical parameter. Since the manual construction of these cartograms is a difficult and time-consuming task, the study of automated construction algorithms became of considerable interest with the increasing computational power of computers.

Most cartogram algorithms compute value-by-area cartograms of the world at a particular abstraction level, e.g., the distribution of wealth among nations in world cartograms. The most common use of these value-by-area world cartograms is to communicate interesting geographic distributions to the public, e.g., worldmapper [2]. Since today's business and scientific applications produce geo-spatial data describing social and economic phenomena of our world at many geographical abstraction levels, there is an increasing need for human analysts to compare the relative importance of regions of interest in the context of world proportions.

As a first promising research step toward interactive geo-spatial exploration systems, we implemented a multi-scale rectangular map approximations approach showing social

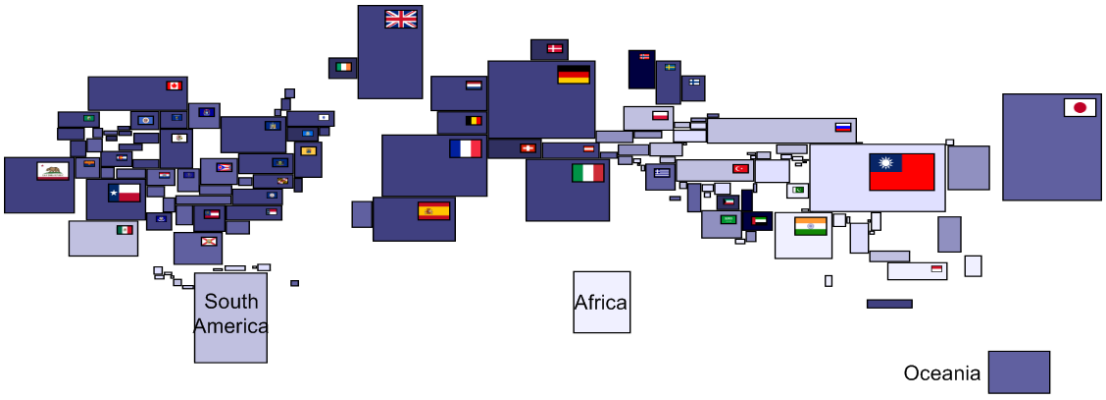


Figure 30.15: Multi-scale rectangular map approximation showing the gross domestic product (gdp) of the world. The size of the regions is proportionate to its gdp. The color of the regions visualizes the gdp per capita: dark blues indicates a very high gdp per capita and a paler color represents a less high gdp. One can easily see that the US federal state California is within the top 20 of the world.

and economic data of the world at different levels of abstractions. In a rectangular map approximation, continents, nations, federal states, etc. are transformed into rectangles, and the displayed area of these rectangles is always proportionate to a given geo-spatial attribute. Multi-scale rectangular map approximations facilitate a better understanding of world relationships by showing detailed views of various regions of interest in the context of world proportions. Figure 30.15 shows the geographical distribution of the gross domestic product across multiple abstraction levels.

We also investigated an integrated approach combining automated analysis with maps to highlight the spatio-temporal behavior of dynamically changing data [3]. The goal is to support analytical reasoning and visual assistance to facilitate good decisions.

References

[1] B. D. Dent, J. S. Torguson, and T. W. Hodler. *Cartography: Thematic Map Design*. Mcgraw Hill, sixth edition, 2008.

[2] D. Dorling, A. Barford, and M. Newman. Worldmapper: The world as you've never seen it before. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):757–764, 2006. Project Website: http://www.worldmapper.org/.

- [3] M. Sips, J. Schneidewind, and D. Keim. Highlighting space-time pattern: Effective visual encodings for interactive decision making. *International Journal of Geographical Information Science (IJGIS)*, 21(8):879–894, 2007.

[4] T. A. Slocum, R. B. McMaster, and F. C. Kessler. *Thematic Cartography and Geovisualization*. Prentice Hall Series in Geographic Information Science, Prentice Hall, third edition, 2008.

[5] S. Steinberg. The world as seen from new york's 9th avenue (cover of the New Yorker march 1976), 1976.

[6] D. K. Wallingford. A new yorker's idea of the united states of america (columbia university press, new york), 1937.

30.6 Integrative Scientific Computing

Coordinator: Robert Strzodka

Beside theory and experiment, computational science established itself as a third valuable mode of scientific investigation. We gain more and more insights from very large and detailed simulations of natural phenomena. Our research focuses on significant improvements of performance and accuracy in scientific computing through a global optimization across the entire spectrum of continuous modeling, numerical analysis, algorithm design, software implementation and hardware acceleration. Currently we concentrate on data layout and solver design for complex domains that exploit the great performance of parallel coprocessors like GPUs, Larrabee or Cell BE without exposing their restrictions and peculiarities to the application programmer.

30.6.1 Mixed Precision Methods

Investigator: Robert Strzodka

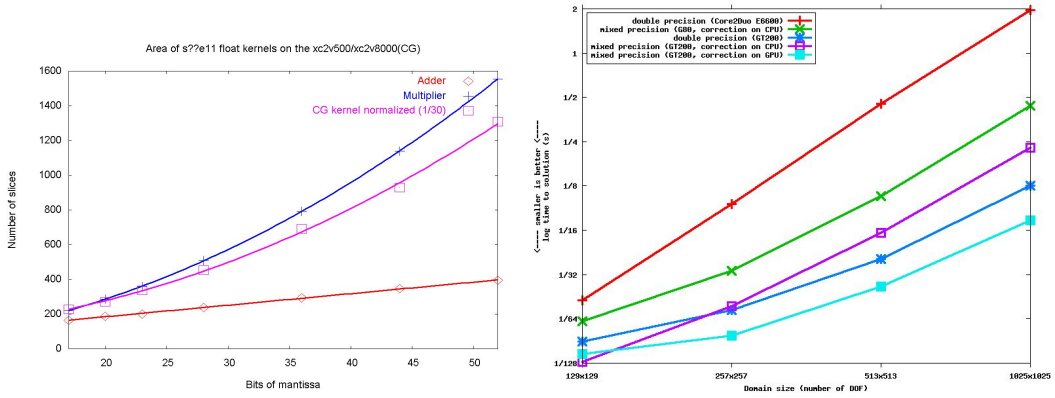


Figure 30.16: On the left: Linear area growth of an adder, the quadratic of a multiplier, and the slower quadratic growth of a conjugate gradient solver in an FPGA. On the right: Comparison of standard double precision and mixed precision multigrid solvers on the CPU and different GPUs. On the older G80 GPU the double correction had to be performed on the CPU, the newer GT200 can do this itself. Note, the logarithmic scale, so that on a 1025x1025 domain the is almost a factor of two between each of the GPU solutions.

Computational precision and the accuracy of the final result have a complicated, non-monotonic relation, so that in general an increase of precision can lead to a decrease of accuracy. While this deteriorating effect of increased precision occurs rather seldom in practice, the non-linear relation often leads to the situation where the accuracy improves differently depending

on which part of the algorithm uses higher computational precision. Numerical analysis methods provide knowledge about these relations, and this can be exploited by reducing the computational precision in less sensitive parts of the algorithm and increasing it in the critical ones. This leads to a *mixed precision method*, which utilizes different computational precision for different parts of the algorithm.

The above idea is particularly attractive for parallel devices with very high single floating point performance, e.g. GPUs, Larrabee, Cell BE, FPGAs. The low precision computations can be executed very quickly in parallel, while the fewer high precision parts of the algorithm that are necessary to obtain the desired accuracy, may be computed in a slower mode on the same device, or can be delegated to the host CPU.

Mixed precision methods benefit both memory bound and computation bound algorithms. While a hardware adder grows linearly in size with the operands, a multiplier grows quadratically, see Figure 30.16 on the left. This means that four 32-bit integer multipliers occupy the same area as one 64-bit integer multiplier, and thus the savings for computation bound algorithms are quadratic in the reductions of the operand size. Further benefits come from the fact that with smaller operands more values can be stored on-chip for faster access.

We have performed a detailed comparison of sparse iterative solvers (conjugate gradient and multigrid) with native-, emulated- and mixed-precision [1]. Figure 30.16 on the left shows results for the mixed precision multigrid solver on the GPU.

References

- [1] D. Göddeke, R. Strzodka, and S. Turek. Performance and accuracy of hardware-oriented native-, emulated- and mixed-precision solvers in FEM simulations. *International Journal of Parallel, Emergent and Distributed Systems*, 22(4):221–256, 2007.

30.6.2 Scientific Computing on GPU-Clusters

Investigator: Robert Strzodka

Commodity based clusters dominate the Top500 supercomputer list in the number of deployed systems. The mass production of commodity components offers advantages in acquisition costs, availability and extensibility, modularity and compatibility. However, these metrics do not tell the whole story. The inefficiency of today's CPUs for high performance computing (HPC) tasks in general and data dominated problems in particular leads to low performance and high power consumption per node, requiring many nodes for large problems and thus additional infrastructure for accommodating, connecting and cooling that many components. In addition, very high numbers of nodes cause the administration and maintenance costs to increase super-linearly.

Parallel add-on coprocessors hold the promise of reducing space requirements and improving the important metrics of performance per cost and performance per power. We have examined this situation by comparing four configurations: 8 original cluster nodes, 16 original cluster nodes, 8 more powerful cluster nodes, and the 8 original cluster nodes enhanced by GPUs [2]. The GPU configuration clearly outperforms the other setting according to the above metrics. The methods discussed in the previous Section 30.6.1 ensure that the accuracy of the GPU configuration is not worse. In the following we have examined the scalability of this

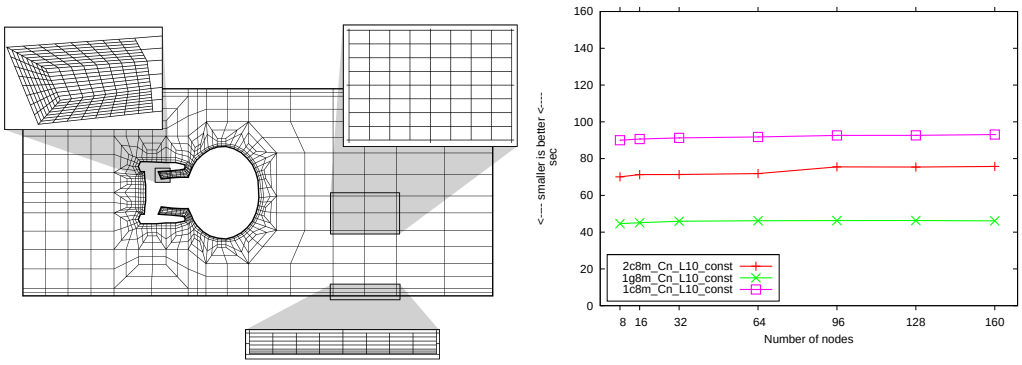


Figure 30.17: On the left: Domain representation in the FE-package FEAST. On the coarse level the subdomains are unstructured, but each subdomain itself is structured allowing faster local processing. On the right: Almost perfect weak scalability results on a cluster with up to 160 nodes in a one CPU and two CPU configuration, and the GPU configuration.

coprocessor acceleration approach by solving test cases on up to 160 GPU enhanced cluster nodes [1]. Figure 30.17 on the right shows the superior GPU performance and the very good weak scalability behavior, although only outdated GPUs were available in such high numbers.

Another important aspect of large scale simulations is the code complexity. On clusters the message passing interface (MPI) is typically used to communicate information among the nodes. This type of low level programming requires careful synchronization and typically application domain specific libraries offer a more high level access to this functionality. We used the FEAST package developed at the University of Dortmund. However, even with the library support the code development of large applications is hard, error prone and time consuming. Therefore, application programmers are very reluctant to rewrite large parts of the code even in view of high performance gains. Bearing this observation in mind, we have developed a *minimally invasive* integration of the hardware acceleration into the FEAST package that does not require any code changes to the applications based on FEAST [2]. Moreover, the changes to the library itself affected also only 1% of its code basis. Although FEAST has not been developed with GPUs in mind, the data centric domain discretization used in the package (Figure 30.17 on the left) was a key factor in achieving such efficient integration.

References

- [1] D. Göddeke, R. Strzodka, J. Mohd-Yusof, P. McCormick, S. H. Buijssen, M. Grajewski, and S. Turek. Exploring weak scalability for FEM calculations on a GPU-enhanced cluster. *Parallel Computing*, 33(10-11):685–699, 2007.
- [2] D. Göddeke, R. Strzodka, J. Mohd-Yusof, P. McCormick, H. Wobker, C. Becker, and S. Turek. Using GPUs to improve multigrid solver performance on a cluster. *International Journal of Computational Science and Engineering*, 4(1):36–55, 2008.

30.6.3 Coprocessor Interoperability: GPU, Larrabee, Cell

Investigators: Zhao Dong and Robert Strzodka

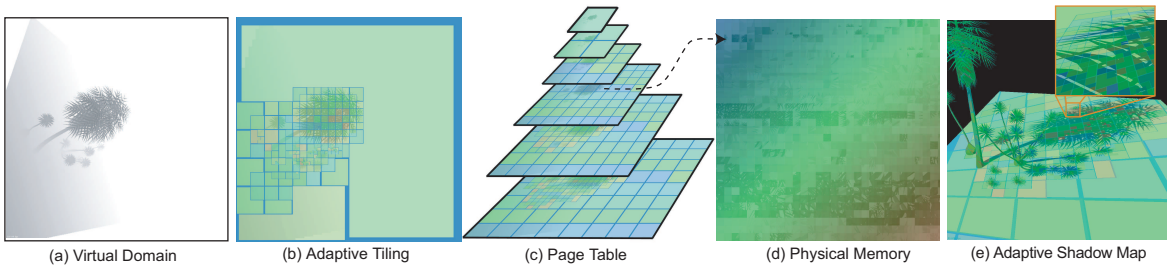


Figure 30.18: Glift offers an abstraction of physical data layout on the GPU. The programmer may use usual coordinates into the virtual (application) domain (a). This domain is adaptively tiled to the resolution where required (b). The data access is performed by first converting the virtual address to a physical address with an address translator (c). This address translator is a minor modification of a generic page table translator that can be used to define many other data structures. The data are then retrieved from the physical memory texture (d). The last image (e) shows an interactive rendering with the adaptive tiles colored for visualization of the different resolutions.

The coprocessors GPU, Larrabee and Cell have been all developed for multimedia work loads and share many common feature on the hardware level but have quite different application programming interfaces (API) for access to their functionality. This is unfortunate as despite their initial application area, they are very well suited as add-on scientific coprocessors even for very large simulations as discussed in the previous Section 30.6.2.

While OpenCL has been developed as a common language for such hardware accelerators, this is a very low level common denominator. On top of the low level languages, we aim at providing a common interface that will enable the unified processing of complex data structures. A predecessor to this work was the Glift project [1] developed jointly with the University of California Davis and the University of Salt Lake City (Figure 30.18). Glift offers parallel access to adaptively refined data structures while separating algorithm design and data layout. Now that the coprocessors offer dedicated on-chip memory this approach can be pushed much further with dynamic work scheduling and faster and more flexible interaction within and across the data blocks.

References

[1] A. E. Lefohn, J. Kniss, R. Strzodka, S. Sengupta, and J. D. Owens. Glift: An abstraction for generic, efficient GPU data structures. *ACM Transactions on Graphics*, 25(1):1–37, 2006.

30.6.4 Bandwidth Reduction Techniques

Investigators: Mohammed Shaheen and Robert Strzodka

The speedups obtained on scientific problems in Section 30.6.2 comes mainly from the superior bandwidth performance of the added coprocessors and far less from the additional parallelism. This is true for many scientific problems, namely that the data transport and not computation is the main performance bottleneck. The development of the hardware interoperable library discussed in the previous Section 30.6.3 also pays close attention to the necessity of bandwidth reduction techniques.

A central technique in this field is the asymmetric traversal of the space-time spanned by the spatial domain of the problem and the time axis representing the multiple iterations to be performed. The naive approach is to advance with the entire spatial domain a single time-step one by one. If the entire spatial domain does not fit the cache, which in practice is almost always true, this naive version incurs many cache misses, cf. green graph in Figure 30.19). A more clever method exploiting data locality is to advance certain parts of the spatial domain multiple time steps at once. One must however respect the data dependencies of the computation between the time steps, such that the computed regions in the space-time are not boxes but pyramids or more generally prismoids. We have investigated different methods for these optimized space-time traversals, obtaining enormous gains over the naive scheme, see Figure 30.19).

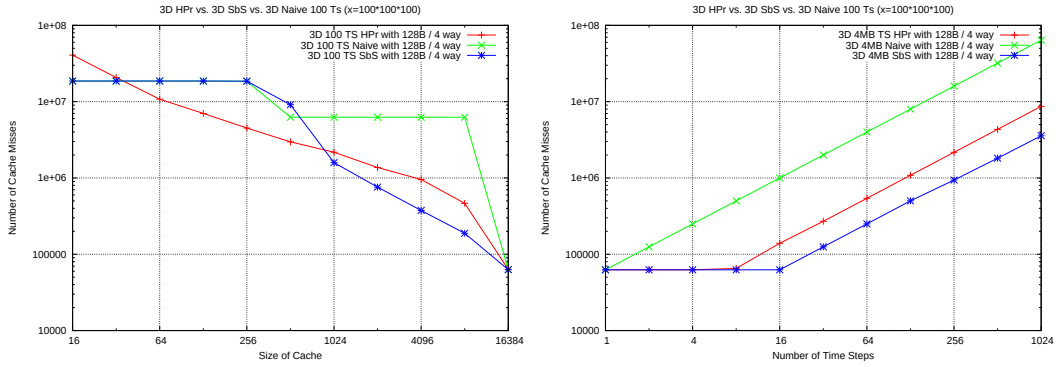


Figure 30.19: Comparison of a naive, hierarchical cache oblivious (HPr), and an aggressive cache aware (SbS) iterative 3D stencil scheme in the number of cache misses. We see on the logarithmic scale that orders of magnitude savings are possible, which is crucial for the performance of data bound algorithms. On the left: Results for 100 time steps and different cache sizes. On the right: Results for 4MiB cache and different number of time steps.

30.6.5 Parallel Coupling of Grids and Particles

Investigators: Mikael Kalms and Robert Strzodka

The coupling of grid and particle information allows to discretize a dynamic simulation more accurately. However, the parallel treatment of this coupling can be challenging because of the different preferred data layouts for grids and particles in view of adaptive nesting (Section 30.6.3) and bandwidth reduction (Section 30.6.4).

In case of free surfaces level set methods are powerful tools to capture their evolution. While the implicit grid based representation deals very well with topological changes, high resolutions and high order integration schemes are required to preserve fine surface details. The particle level set (PLS) method is an important extension which additionally employs particles to correct the numerical dissipation in the grid, see Figure 30.20 on the left. This allows to re- duce the spatial resolution and the order of the integration, while preserving a time consistent evolution. We have developed an enhanced variant of the PLS approach that achieves both, higher performance and superior quality in terms of mass preservation [1].

A second project deals with the accurate simulation of large cosmological constellations. These are represented mainly as particle systems, but grid based data comes into play through the continuous representation of gas clouds, see Figure 30.20 on the right. In a certain sense this is a dual representation problem in comparison to the free surfaces. Now the particles are the dominating structure and the grid is the auxiliary construct.

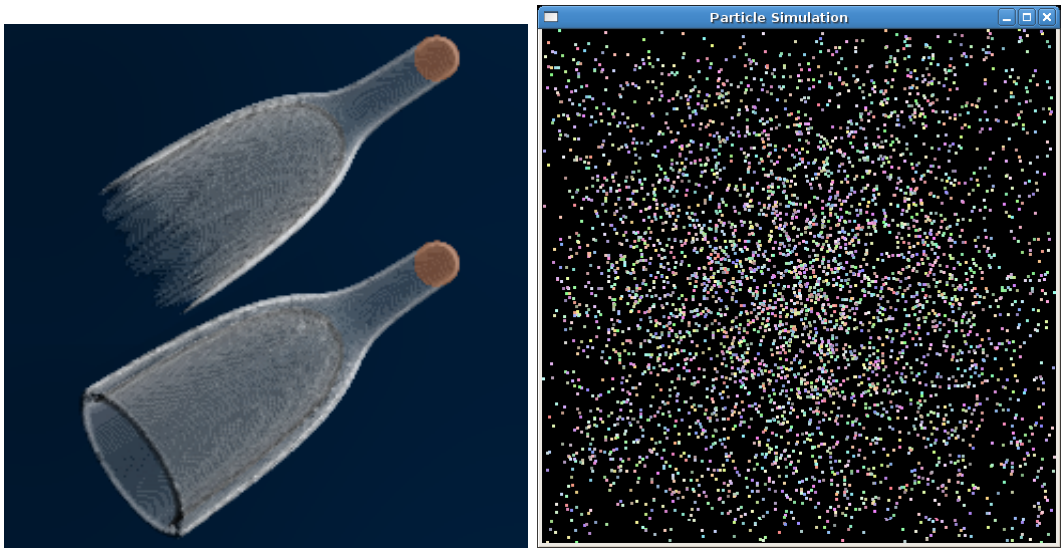


Figure 30.20: On the left: Comparison of a free surface evolution with grid based level set representation alone, and a hybrid particle grid representation. On the right: Simulation of a large particle system, with the requirement to couple particles with continuously distributed matter.

References

- [1] N. Cuntz, A. Kolb, R. Strzodka, and D. Weiskopf. Particle level set advection for the interactive visualization of unsteady 3D flow. *Computer Graphics Forum (Proc. EuroVis)*, 27(3):719–726, 2008.

30.7 Markerless Motion Capture and Multiview Stereo Processing

Coordinators: Bodo Rosenhahn and Thorsten Thormählen

This area of research deals with the problem of estimating the motion and shape of articulated objects as well as the static 3D scene background and the camera parameters from image sequences or video. The novel algorithms and mathematical representations developed during this reporting period have let to a new quality of reconstruction, both in terms of robustness and accuracy.

30.7.1 Camera Motion Estimation and Static Scene Reconstruction

Investigators: Nils Hasler and Thorsten Thormählen

For automatic camera motion tracking, a unified technique for merging unconnected feature tracks was introduced [2], which allows fully automatic camera motion estimation in difficult situations. In particular, we investigated three scenarios: drift removal, registration of multiple moving cameras (see Fig. 30.21), and recovery of camera motion after large occlusions.

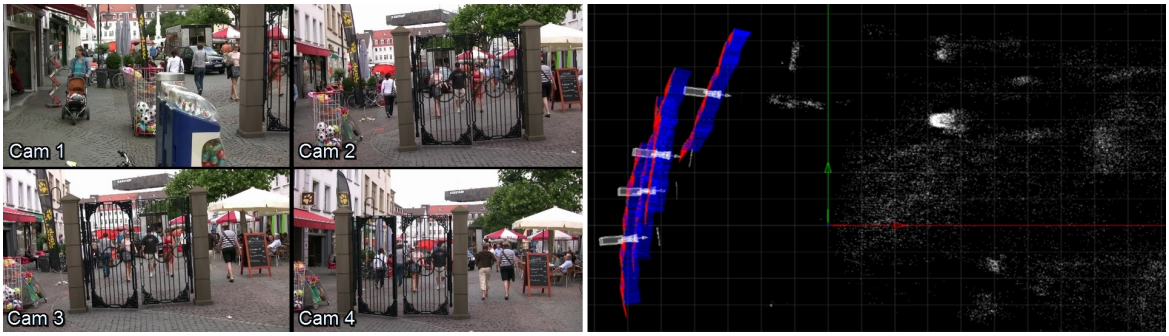


Figure 30.21: Registration of multiple moving cameras: A sample image out of the sequence for each of the four cameras is shown (left). The images are augmented with a 3D model of a gate. Top view on the estimated camera path of the four moving cameras (right).

For reconstruction of a static 3D scene, a semi-automatic approach was developed [3] that enables the generation of a high-quality 3D model of a static object from an image sequence that was taken by a moving, uncalibrated consumer camera. First, the camera parameters for each input image are estimated by automatic camera tracking. Afterwards, approaches from image-based rendering are used to generate an orthographic projection on a bounding box that is placed around the object. These ortho-images can be imported as

background maps in the orthographic views of any modeling package (e.g., the top, side, and front view). Now, modelers can use the ortho-images to guide their modeling with the familiar tools of their modeling package. Thereby, they can use all the advanced features that the modeling package has to offer, such as spline modeling or subdivision methods. Because of the orthographic projection, the 3D information is directly given by the 2D user interactions in the orthographic views. This greatly improves the accuracy and efficiency of the manual modeling process. The approach is capable of handling not only diffuse surfaces but even translucent or specular surfaces and is therefore still applicable where today's laser scanners or fully automatic image-based approaches would generate inaccurate results.

Furthermore, the recall rate of the well-known SIFT-descriptor was improved [1] by an irregular sampling grid.

References

- [1] Y. Cui, N. Hasler, T. Thormählen, and H.-P. Seidel. Scale invariant feature transform with irregular orientation histogram binning. In *International Conference on Image Analysis and Recognition (ICIAR 2009)*, Halifax, Canada, 2009, Lecture Notes in Computer Science. Springer. (to appear).
- [2] T. Thormählen, N. Hasler, M. Wand, and H.-P. Seidel. Merging of feature tracks for camera motion estimation from video. In *5th European Conference on Visual Media Production (CVMP 2008)*, London, UK, 2008. The Institution of Engineering and Technology.
- [3] T. Thormählen and H.-P. Seidel. 3d-modeling by ortho-image generation from image sequences. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 27(3):86:1–86:5, 2008.

30.7.2 Markerless Motion Capture

Investigators: Nils Hasler, Jürgen Gall, Bodo Rosenhahn, Martin Sunkel, and Thorsten Thormählen

Typical cues for motion capture are silhouettes, edges, color, motion, and texture. In general, a multi-cue integration is necessary for tracking complex objects, like humans, since all these cues come along with inherent drawbacks. Ideally, the impact of a cue should be large in situations when its extraction is reliable, and small, if the information is likely to be erroneous. To this end, we propose an adaptive weighting scheme that combines complementary cues, namely silhouettes on one side and optical flow as well as local descriptors on the other side [3]. Relying only on image features that are tracked over time does not prevent the accumulation of small errors which results in a drift away from the target object. The error accumulation becomes even more problematic in the case of multiple moving objects due to occlusions. To solve the drift problem for tracking, we propose an analysis-by-synthesis framework that uses reference images to correct the pose. It comprises an occlusion handling and is successfully applied to crash test video analysis [9, 12, 7].

In order to overcome the drawbacks of local optimization, we introduce a novel global stochastic optimization approach for markerless human motion capturing that is derived from the mathematical theory on interacting particle systems [10]. It estimates the human pose without initial information, which is a challenging optimization problem in a high dimensional space [8]. Furthermore, we propose a tracking framework that is based on this optimization

technique to achieve both the robustness of filtering strategies and a remarkable accuracy. In order to benefit from optimization and filtering, we introduce a multi-layer framework that combines stochastic optimization, filtering, and local optimization [6]. While the first layer relies on interacting simulated annealing, the second layer refines the estimates by filtering and local optimization such that the accuracy is increased and ambiguities are resolved over time without imposing restrictions on the dynamics. In addition, we propose a system that recovers not only the movement of the skeleton, but also the possibly non-rigid temporal deformation of the 3D surface [11]. Finally, we have developed a method for detecting pedestrians in still images [5].

Additional information useful in markerless motion capture is encoded in the body shape of the subject. For most methods a priori information about the body shape is necessary. We have investigated a means to drop this requirement by creating a statistical model of human body shapes. Although this last step has not been made yet, the foundation has been laid by creating the model and applying it to body shape estimation of dressed humans from 3D scans [15, 16].

Exceedingly elaborate prior knowledge about the expected motion was employed to stabilize tracking [2, 27, 21, 24, 1, 22, 4, 26, 13, 20, 17, 28, 23].

Additionally, efforts were made to increase the simplicity of the experimental setup for markerless motion capture experiments [14]. Furthermore, tools for human gesture modeling and animation were developed [18, 19] and an improved recognition algorithm for footprints of cryptic species was introduced [25].

References

- [1] T. Brox, B. Rosenhahn, and D. Cremers. Contours, optic flow, and prior knowledge: Cues for capturing 3D human motion in videos. In B. Rosenhahn, R. Klette, and D. Metaxas, eds., *Human Motion - Modeling, Tracking, Capture and Animation*, *Computational Imaging and Vision*, vol. 36, ch. 11, pp. 265–293. Springer, Dordrecht, The Netherlands, 2008.
- [2] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. Nonparametric density estimation with adaptive anisotropic kernels for human motion tracking. In A. Elgammal, B. Rosenhahn, and R. Klette, eds., *2nd Workshop on Human Motion*, Rio de Janeiro, Brazil, 2007, *LNCS 4814*, pp. 152–165. Springer.
- [3] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers. Combined region- and motion-based 3d tracking of rigid and articulated objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, X, 2009.
- [4] A. Elgammal, B. Rosenhahn, and R. Klette, eds. *2nd Workshop: Human Motion - Understanding, Modeling, Capture and Animation*, *LNCS 4814*. Springer, Berlin, Germany, 2007.
- [5] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, Miami, USA, 2009, pp. 1–8. IEEE Computer Society.
- [6] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture - a multi-layer framework. *International Journal of Computer Vision*, pp. 1–18, 2009. To appear.
- [7] J. Gall, B. Rosenhahn, S. Gehrig, and H.-P. Seidel. Model-based motion capture for crash test video analysis. In G. Rigoll, ed., *Pattern Recognition*, Munich, Germany, 2008, *LNCS 5096*, pp. 92–101. Springer.

Figure 30.22: An example of a motion capture result generated using moving, automatically synchronized cameras (cf. [14]).

- [8] J. Gall, B. Rosenhahn, and H.-P. Seidel. Clustered stochastic optimization for object recognition and pose estimation. In L. Hamprecht, C. Schnörr, and B. Jähne, eds., *29th Annual Symposium of the German Association for Pattern Recognition (DAGM'07)*, Heidelberg, Germany, 2007, *LNCS 4713*, pp. 32–41. Springer.
- [9] J. Gall, B. Rosenhahn, and H.-P. Seidel. Drift-free tracking of rigid and articulated objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, USA, 2008, pp. 1–8. IEEE Computer Society.
- [10] J. Gall, B. Rosenhahn, and H.-P. Seidel. An introduction to interacting simulated annealing. In R. Klette, D. Metaxas, and B. Rosenhahn, eds., *Human Motion - Understanding, Modeling, Capture and Animation*, pp. 319–345. Springer, Heidelberg, 2008.
- [11] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, Miami, USA, 2009, pp. 1–8. IEEE Computer Society.
- [12] S. Gehrig, B. Hernán, and J. Gall. Accurate and model-free pose estimation of crash test dummies. In R. Klette, D. Metaxas, and B. Rosenhahn, eds., *Human Motion - Understanding, Modeling, Capture and Animation*, pp. 453–473. Springer, Heidelberg, 2008.

- [13] D. Han, B. Rosenhahn, S. Gehrig, and H.-P. Seidel. Combined registration methods for pose estimation. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, F. M. Porikli, J. Peters, J. Klosowski, L. Arns, Y. Chun, T.-M. Rhyne, and L. Monroe, eds., *Proceedings of the 4th International Symposium on Advances in Visual Computing (ISVC 2008)*, Las Vegas, NV, USA, 2008, *LNCS 5358*, pp. 913–924. Springer.
- [14] N. Hasler, B. Rosenhahn, T. Thormählen, M. Wand, J. Gall, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, USA, 2009. IEEE Computer Society. (to appear).
- [15] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Munich, Germany, 2009, vol. 2, pp. 337–346. Blackwell.
- [16] N. Hasler, C. Stoll, T. Thormählen, B. Rosenhahn, and H.-P. Seidel. Estimating body shape of dressed humans. *Computers & Graphics*, 2009. (to appear).
- [17] M. Müller, B. Demuth, and B. Rosenhahn. An evolutionary approach for learning motion class patterns. In G. Rigoll, ed., *Pattern Recognition*, Munich, Germany, 2008, *LNCS 5096*, pp. 365–374. Springer.
- [18] M. Neff, I. Albrecht, and H.-P. Seidel. Layered performance animation with correlation maps. In D. Cohen-Or and P. Slavik, eds., *Eurographics 2007*, Prague, Czech Republic, 2007, *Computer Graphics Forum*, vol. 26, pp. 675–684. Blackwell.
- [19] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics*, 27(1):5, 2008.
- [20] B. Rosenhahn, T. Brox, D. Cremers, and H.-P. Seidel. Modeling and tracking line-contrained mechanical systems. In G. Sommer and R. Klette, eds., *Robot Vision, 2nd International Workshop*, Auckland, New Zealand, 2008, *LNCS 4931*, pp. 98–110. Springer.
- [21] B. Rosenhahn, T. Brox, and H.-P. Seidel. Scaled motion dynamics for markerless motion capture. In *2007 IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07. - Vol. 3*, Minneapolis, Minnesota, 2007, pp. 1203–1210. IEEE.
- [22] B. Rosenhahn, U. Kersting, K. Powell, T. Brox, and H.-P. Seidel. Tracking clothed people. In B. Rosenhahn, R. Klette, and D. Metaxas, eds., *Human Motion Understanding,. Modeling, Capture and Animation*, *Computational Imaging and Vision*, vol. 36, ch. 12, pp. 295–317. Springer, Dordrecht, The Netherlands, 2008.
- [23] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, D. Cremers, and H.-P. Seidel. Markerless motion capture of man-machine interaction. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.
- [24] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, and H.-P. Seidel. Staying well grounded in markerless motion capture. In G. Rigoll, ed., *Proceedings of the 30th DAGM Symposium*, München, Germany, 2008, *LNCS 5096*, pp. 385–395. Springer.
- [25] J. Russell, N. Hasler, R. Klette, and B. Rosenhahn. Automatic track recognition of footprints for identifying cryptic species. *Ecology*, 2009. (to appear).
- [26] C. Schmaltz, B. Rosenhahn, T. Brox, D. Cremers, J. Weickert, L. Wietzke, and G. Sommer. Region-based pose tracking. In J. Marti, J.-M. Benedi, A. Mendonca, and J. Serrat, eds., *Proceedings of the 3rd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2007)*, Girona, Spain, 2007, *LNCS 4478*, pp. 56–63. Springer.

- [27] C. Schmaltz, B. Rosenhahn, T. Brox, J. Weickert, L. Wietzke, and G. Sommer. Dealing with self-occlusion in region based motion capture by means of internal regions. In F. J. López and R. B. Fisher, eds., *Proceedings of the 5th International Conference on Articulated Motion and Deformable Objects (AMDO 2008)*, Port d'Andratx, Mallorca, Spain, 2008, *LNCS 5098*, pp. 102–111. Springer.

- [28] M. Sunkel, B. Rosenhahn, and H.-P. Seidel. Silhouette based generic model adaptation for marker-less motion capturing. In A. Elgammal, B. Rosenhahn, and R. Klette, eds., *Human Motion - Understanding, Modeling, Capture and Animation, Second Workshop, Human Motion 2007*, Rio de Janeiro, Brazil, 2007, *LNCS 4814*, pp. 119–135. Springer.

30.7.3 Performance Capture and Virtual Actors

Investigators: Edilson de Aguiar, Naveed Ahmed, and Carsten Stoll

Nowadays, stepping directly from a captured real-world sequence (Fig. 30.23(a)) to the corresponding realistic moving character (Fig. 30.23(b,c)) is still challenging. Since marker-based and marker-free motion capture systems measure the motion in terms of a kinematic skeleton, they have to be combined with other scanning technologies to capture the time-varying shape and surface details of the human body surface. However, dealing with people wearing arbitrary clothing from only video streams is still not possible.

To overcome this limitation, we researched a new framework [1, 8, 2], enabling the direct animation of a high-quality static human scan from unaltered video footage. Our algorithms jointly capture motion and time-varying shape detail even of people wearing wide apparel, while preserving its spatio-temporally coherence over time. By being completely passive, they also enable us to record the subject's appearance, which can then be used to display the recorded actor from arbitrary viewpoints.

The proposed methods achieve a high level of flexibility and versatility by explicitly abandoning any traditional skeletal or motion parameterization, and by posing *performance capture* as deformation capture. As a result, they produce a rich dynamic scene representation that can be easily made available to and modified by animators, which enables new applications in motion capture, computer animation and 3D Video [6].

While performance capture methods can already capture high-quality representations of actors, the resolution of features which they can detect is still limited. By extending on ideas presented in our work on Relightable Free-viewpoint video [9], we are able to reconstruct a time-varying normal field on top of our geometry which allows us to reconstruct surface features at millimeter scale, as for example small folds and wrinkles [4] (Fig. 30.24 (a-d)).

Novel mesh deformation methods such as [7] as well as new surface based scene capture techniques offer a great level of flexibility during animation creation. However, unfortunately the resulting scene representation is less compact than skeletal ones and there is not yet a rich toolbox available which enables easy post-processing and modification of mesh animations.

Animators are used to a large repertoire of tools for editing and rendering traditional skeletal animations, but yet lack the same set of tools for working with mesh-based dynamic scene representations. Our method proposed in [3] bridges this gap, enabling the fully-automatic conversion of a mesh animation into a skeleton-based animation, Fig. 30.23(d,e).

A second novel application is described in [10]. Usually, researchers in computer graphics aim at developing algorithms that enable the computer and even unexperienced users to

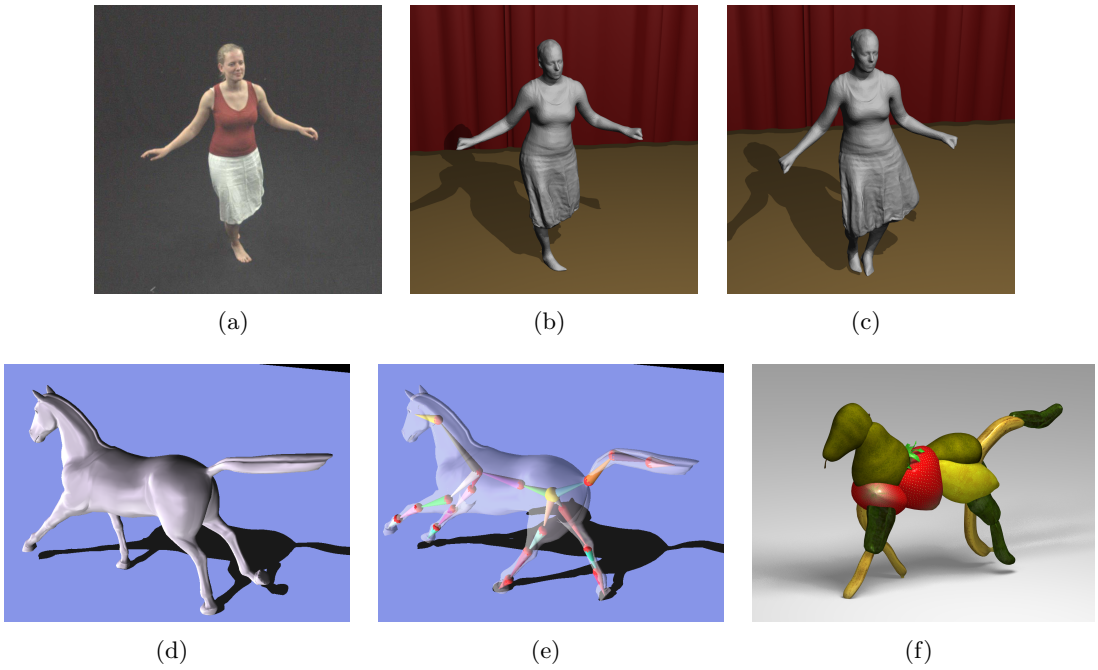


Figure 30.23: Performance Capture: (a) input real-world sequence and two reconstructions where the lifelike folds and wrinkles visible in the input have been captured (b,c). Processing Mesh Animations: (d) Input mesh animation and (e) automatically generated skeleton-based animation; (f) resulting animation collage using a database of fruits.

reproduce the look of certain styles of visual arts. Our system brings together the traditional art form of a collage with the most prominent art form in computer graphics, namely 3D animation, allowing the computer artist to automatically convert her favorite mesh animation into a moving assembly of 3D shape primitives from a database, Fig. 30.23(f).

Creating spatio-temporally coherent mesh animations from other scene representations has been the main goal pursued in [5]. Using shape-from-silhouette methods allows to reconstruct detailed surfaces for each timestep of a video but yield sequences of shapes made of triangle meshes with varying connectivity. By first establishing sparse 3D correspondences from image features and then extending these to densely cover the whole mesh we can re-parameterize the sequence to generate a single spatio-temporally coherent animation (Fig. 30.24 (e,f)).

References

- [1] E. de Aguiar, C. Stoll, N. Ahmed, and H.-P. Seidel. Performance capture from sparse multi-view video. In G. Turk, ed., *Proceedings of ACM SIGGRAPH 2008*, Los Angeles, USA, 2008, *ACM Transactions on Graphics*, vol. 27(3), pp. 1–10. ACM.

- [2] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less 3d feature tracking for mesh-based motion capture. In A. Elgammal, B. Rosenhahn, and R. Klette, eds., *Human Motion* -

(a) (b) (c) (d) (e) (f)

Figure 30.24: Fine geometry reconstruction: (a,c) input frames with corresponding reconstructed surface detail rendered from same camera viewpoint (b,d). Correspondence finding: (e,f) spatio-temporally coherent meshes generated from input incoherent shape-from-silhouette sequence.

Understanding, Modeling, Capture and Animation, Rio de Janeiro, Brazil, 2007, *LNCS 4814*, pp. 1–15. Springer.

- [3] E. de Aguiar, C. Theobalt, S. Thrun, and H.-P. Seidel. Automatic conversion of mesh animations into skeleton-based animations. In R. Scopigno and E. Gröller, eds., *Computer Graphics Forum (Proceedings Eurographics EG'08)*, Hersonissos, Crete, Greece, 2008, vol. 27(2), pp. 389–397. Blackwell.
- [4] N. Ahmed, C. Theobalt, P. Dobrev, H.-P. Seidel, and S. Thrun. Robust fusion of dynamic shape and normal capture for high-quality reconstruction of time-varying geometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage, Alaska, 2008, pp. 1–8. IEEE Computer Society.
- [5] N. Ahmed, C. Theobalt, C. Rössl, S. Thrun, and H.-P. Seidel. Dense correspondence finding for parametrization-free animation reconstruction from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage, Alaska, 2008, pp. 1–8. IEEE Computer Society.
- [6] M. Eisemann, B. de Decker, M. A. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt, and A. Sellent. Floating textures. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 27(2):409–418, 2008.
- [7] C. Stoll, E. de Aguiar, C. Theobalt, and H.-P. Seidel. A volumetric approach to interactive shape editing. Research Report MPI-I-2007-4-004, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, 2007.
- [8] C. Theobalt, E. de Aguiar, M. Magnor, and H.-P. Seidel. Reconstructing human shape, motion and appearance from multi-view video. In H. Ozaktas and L. Onural, eds., *Three-Dimensional Television: Capture, Transmission, and Display*, pp. 29–58. Springer, Heidelberg, Germany, 2008.
- [9] C. Theobalt, N. Ahmed, H. P. A. Lensch, M. Magnor, and H.-P. Seidel. Seeing people in different light-joint shape, motion, and reflectance capture. *IEEE Transactions on Visualization and Computer Graphics*, 13(4):663–674, 2007.
- [10] C. Theobalt, C. Rössl, E. de Aguiar, and H.-P. Seidel. Animation collage. In M. Gleicher and D. Thalmann, eds., *Symposium on Computer Animation*, San Diego, California, 2007, pp. 271–280. Eurographics.

30.7.4 Faces

Investigators: Robert Bargmann, Volker Blanz, and Kristina Scherbaum

a) b) c) d)

Figure 30.25: The generated textures (b) and (d) are generated from the scans in the left columns (a) and (c). The inversion of illumination effects has removed most of the harsh lighting from the original textures. The method compensates both the results of overexposure and inhomogeneous shading of the face.

Facial animation and manipulation in movie productions has reached an impressive level of realism today. In particular, the shapes designed by animators are highly realistic and expressive. However, this manual design is extremely labor-intensive. In order to simplify and accelerate this process, a number of automated techniques have been proposed in computer graphics. In learning-based approaches, a knowledge database must be build once, while the synthesis is produced very quickly. Fast scanning techniques, which are becoming more and more available today, provide a unique source of information about the facial tissue and its motions to learn from. The goal of this group is to create a powerful, robust and general approach to analyzing such data. In a consequence, we apply the learned knowledge to synthesize specific properties of facial shapes and to simulate facial articulation. This includes applications that range from face animation to face recognition.

For example, we presented a new method ([1]) that generates synthetic mouth articulations from single audio files, which may be transferred to any other facial mesh of an arbitrary person. The core part of this algorithm is based on learning articulations from a stream of 3D scans of a real person. These scans are acquired by a structured light scanner at 40 reconstructions per second. Correspondences between the scans over several speech sequences are established via optical flow. On the registered mesh a novel type of Principal Component Analysis is computed that considers variances only in a sub-region of the face, while retaining the full dimensionality of the original vector space of sample scans. Audio is recorded at the same time, so the head scans can be synchronized with phoneme and viseme information for computing viseme clusters. Given a new audio sequence along with text data, we are able to automatically create an animation for that new sentence by morphing between the visemes along a path in viseme-space. These methods include an automated process for data analysis in streams of 3D scans, and a framework that connects the system to existing static face modeling technology for articulation transfer.

In another project we developed a new top-down approach to 3D data analysis that

specifically exploits the fact, that most of todays 3D scanners record surfaces in a perspective projection. Due to the rapid progress of such 3D scanning technology face recognition from 3D scans has become a very active research field. In contrast to standard 2D face recognition, multimodal approaches allow for the compensation of changes in pose and illumination. Our goal was to build a framework, that allows to easily compare 2D and 3D recordings of faces. To achieve a uniform representation of 2D and 3D inputs we fit a 3D Morphable Face Model to both, the photographs and the 3D scans of faces. In this context we proposed a new method for the 3D fitting problem: To adapt the Morphable Model to facial 3D meshes, we follow an analysis-by-synthesis approach, where shape, texture, pose and illumination are optimized simultaneously. Starting from raw 3D scans, the algorithm determines a PCA-based representation which fits the scan best. Also, fragmentary surfaces are completed and correspondences to a reference face of the Morphable Model are established. Simultaneously, illumination conditions are estimated in an explicit simulation that involves specular and diffuse components. The effects of lighting and shading are removed to obtain an illumination corrected texture, which stores the diffuse reflectance in each point of the facial surface as shown in Figure 30.25. Finally, we investigated a face recognition scenario with our algorithm on a set of facial recordings (2D and 3D), and evaluated how the additional shape information improves the performance compared to image-based methods [2]. Our results demonstrate that analysis-by-synthesis is not only a promising strategy in image analysis, but can also be applied to range data. We applied a similar 3D fitting technique also in another context, where we predicted the facial growth of children as learned from a database of facial 3D scans [3].

References

- [1] R. Bargmann, V. Blanz, and H.-P. Seidel. A nonlinear viseme model for triphone-based speech synthesis. Research Report MPI-I-2007-4-003, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, 2007.
- [2] V. Blanz, K. Scherbaum, and H.-P. Seidel. Fitting a morphable model to 3d scans of faces. In *Eleventh IEEE International Conference on Computer Vision*, Rio de Janeiro, Brasil, 2007, *DVD Proceedings*, vol. CFP07198-CDR, pp. 1–8. IEEE ICCV 2007, Omnipress.
- [3] K. Scherbaum, M. Sunkel, H.-P. Seidel, and V. Blanz. Prediction of individual non-linear aging trajectories of faces. In D. Cohen-Or and P. Slavik, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Prague, Czech Republic, 2007, vol. 26(3), pp. 285–294. Blackwell.

30.8 Multimedia Information Retrieval

Coordinator: Meinard Müller

Modern information society is experiencing an explosion of digital content, comprising text, audio, video, and graphics. The challenge is to organize, understand, and search multimodal information in a robust, efficient, and intelligent manner. One challenge arises from the fact that multimedia objects, even though they are similar from a structural or semantic viewpoint, often reveal significant spatial or temporal differences. This makes content-based multimedia retrieval a challenging research field with many unsolved problems. The goal of our group is

to develop fundamental algorithms and concepts for the analysis, classification, indexing, and retrieval of time-dependent data streams. In particular, we deal with two different multimedia domains: music data and human motion data.

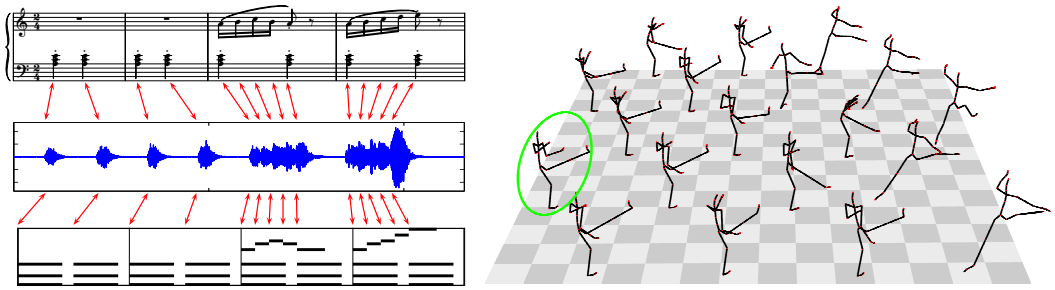


Figure 30.26: **Left:** Synchronization of several types of music data. **Right:** Content-based motion retrieval.

In the music domain, we advance the development of techniques and tools for organizing, structuring, retrieving, navigating, and presenting music-related data. Here, one objective is to automatically link several types of music data including text, symbolic data, audio, image, and video with the goal to coordinate the multiple information sources related to a given musical work. In the motion domain, we explore new approaches to motion analysis, retrieval, and classification. One of our strategies is to handle spatio-temporal motion deformations already on the feature level, which then allows us to adopt efficient indexing methods resulting in flexible and efficient retrieval algorithms that are applicable to large motion capture data sets.

30.8.1 Music Synchronization

Investigators: Meinard Müller, Peter Grosche in cooperation with Sebastian Ewert

The general goal of music synchronization is to automatically align different versions and interpretations related to a given musical work [3]. In computing such alignments, recent approaches assume that the versions to be aligned correspond to each other with respect to their overall global structure. However, in real-world scenarios, this assumption is often violated. For example, for a popular song there often exist various structurally different album, extended, or cover versions. Or, in classical music, different recordings of the same piece may exhibit omissions of repetitions or significant differences in parts such as solo cadenzas. In [4], we present a novel synchronization procedure, which can compute meaningful audio alignments even in the presence of structural variations. Here, as one main contribution, we introduce the concept of path-constrained similarity matrices. This enables us to employ a flexible and efficiently computable partial matching procedure in the optimization step of our synchronization algorithm. Furthermore, in [5], we introduce a novel approach for automatically detecting structural similarities and differences between two given versions of the same piece. Here, the key idea is to perform a single structural analysis for both versions

simultaneously instead of performing two separate analyses for each of the two versions. Such a joint structure analysis reveals the repetitions within and across the two versions.

In computing music alignments, one typically has to face a delicate tradeoff between robustness and accuracy. In this context, chroma-based music features are a well-established tool for analyzing and comparing music data. By identifying spectral components that differ by a musical octave, chroma features show a high degree of invariance to variations in timbre. In [1, 2], we introduce novel audio features that combine the high temporal accuracy of onset features with the robustness of chroma features. We show how previous synchronization methods can be extended to make use of these new features. Furthermore, in [6], we describe a novel procedure for making chroma features even more robust to changes in timbre and instrumentation while keeping their discriminative power. Our idea is based on the generally accepted observation that the lower mel-frequency cepstral coefficients (MFCCs) are closely related to timbre. Now, instead of keeping the lower coefficients, we discard them and only keep the upper coefficients. Furthermore, using a pitch scale instead of a mel scale allows us to project the remaining coefficients onto the twelve chroma bins.

References

- [1] S. Ewert and M. Müller. Refinement strategies for music synchronization. In *Proceedings of the 5th International Symposium on Computer Music Modeling and Retrieval (CMMR 2008)*, Copenhagen, Denmark, 2008, Lecture Notes in Computer Science. Springer. To appear.
- [2] S. Ewert, M. Müller, and P. Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009. IEEE.

[3] M. Müller. *Information Retrieval for Music and Motion*. Springer, 2007.

- [4] M. Müller and D. Appelt. Path-constrained partial music synchronization. In *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, U.S.A., 2008, pp. 65–68. IEEE.
- [5] M. Müller and S. Ewert. Joint structure analysis with applications to music annotation and synchronization. In J. P. Bello, E. Chew, and D. Turnbull, eds., *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, Pennsylvania, USA, 2008, pp. 389–394.
- [6] M. Müller, S. Ewert, and S. Kreuzer. Making chroma features more robust to timbre changes. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009. IEEE.

30.8.2 Music Retrieval and User Interfaces

Investigator: Meinard Müller in cooperation with the Multimedia Signal Processing Group *at Bonn University*

Given a large audio database of music recordings, the goal of classical *audio identification* is to identify a particular audio recording by means of a short audio fragment. Even though recent identification algorithms show a significant degree of robustness towards noise, MP3 compression artifacts, and uniform temporal distortions, the notion of similarity is rather close to the identity. In [5], we address a higher level retrieval problem, which we refer to

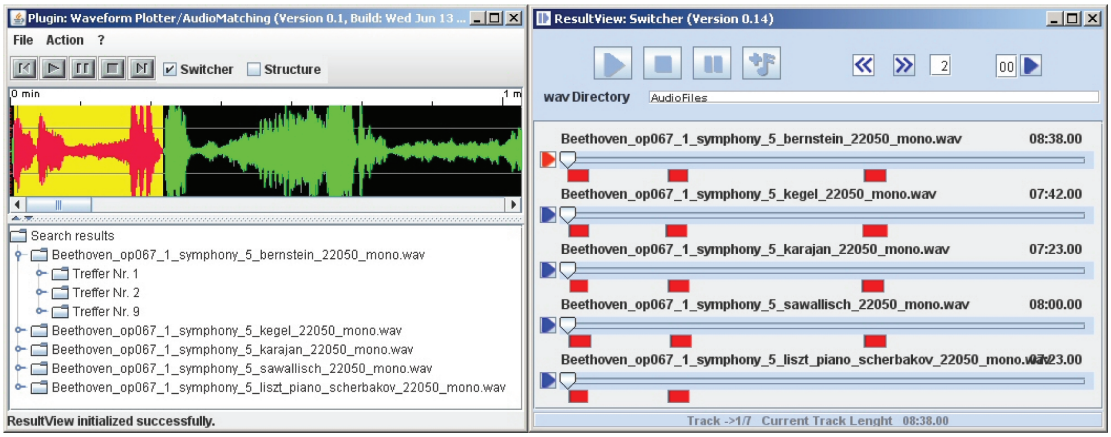


Figure 30.27: Given a short query audio clip (indicated by the yellow segment in the waveform), the goal of audio matching is to automatically retrieve all excerpts from all recordings within the database that musically correspond to the query.

as *audio matching*: given a short query audio clip, the goal is to automatically retrieve all excerpts from all recordings within the database that *musically* correspond to the query, see Fig. 30.27. In our matching scenario, opposed to classical audio identification, we allow semantically motivated variations as they typically occur in different interpretations of a piece of music. To this end, we present an efficient and robust audio matching procedure that works even in the presence of significant variations, such as *non-linear* temporal, dynamical, and spectral deviations, where existing algorithms for audio identification would fail. Furthermore, the combination of various deformation- and fault-tolerance mechanisms allows us to employ standard indexing techniques to obtain an *efficient, index-based* matching procedure, thus providing an important step towards semantically searching large scale real-world music collections.

In [3], we consider the scenario of a multimodal music collection that comprises visual (scanned sheet music) as well as acoustic material (audio recordings). We present a novel procedure for mapping scanned pages of sheet music to a given collection of audio recordings by identifying musically corresponding audio clips. To this end, both the scanned images as well as the audio recordings are first transformed into a common feature representation using optical music recognition (OMR) and methods from digital signal processing, respectively. Based on this common representation, a direct comparison of the two different types of data is facilitated similar to [5]. This allows for a search of scan-based queries in the audio collection. We report on systematic experiments conducted on the corpus of Beethoven's piano sonatas showing that our mapping procedure works with high precision across the two types of music data in the case that there are no severe OMR errors.

To prove the practicability and relevance of our synchronization and retrieval techniques in a real-world applications, we developed a framework for managing heterogeneous digitized music collections [6]. In the last two years, we particularly focused on the multimodal scenario having visual music representations (scanned sheet music) as well as acoustic music

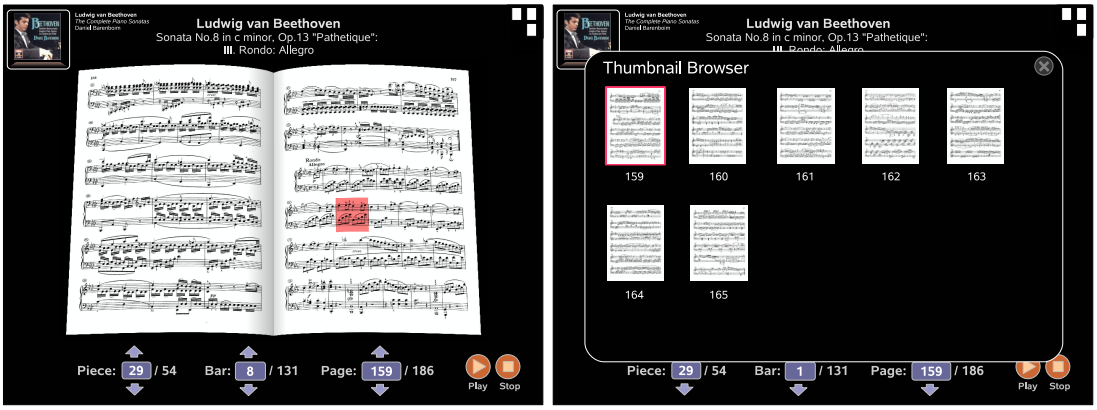


Figure 30.28: The Score Viewer Interface for multimodal music presentation and navigation. Synchronously to audio playback, corresponding musical measures within the sheet music are highlighted (left). The Thumbnail Browser (right) allows to conveniently navigate through the currently selected score.

material (audio recordings). In [4], we propose a preprocessing workflow that comprises feature extraction, audio indexing, and music synchronization linking the visual with the acoustic data. In [1, 2], we introduce novel user interfaces for multimodal music presentation, navigation, and content-based retrieval. In particular, our system offers high quality audio playback with time-synchronous display of the digitized sheet music, see Fig. 30.28. Among others, the listener is enabled to seamlessly crossfade between various interpretations belonging to the currently selected musical work. Furthermore, our system allows a user to select regions within the scanned pages of a musical score in order to search for musically similar sections within the audio documents.

References

- [1] D. Damm, C. Fremerey, F. Kurth, M. Müller, and M. Clausen. Multimodal presentation and browsing of music. In *IMCI '08: Proceedings of the 10th International Conference on Multimodal Interfaces*, Chania, Crete, Greece, 2008, pp. 205–208. ACM.
- [2] D. Damm, C. Fremerey, F. Kurth, M. Müller, and M. Clausen. SyncPlayer - Multimodale Wiedergabe, Navigation und Suche in heterogenen digitalen Musikkollektionen. In T. Mandl, N. Fuhr, and A. Henrich, eds., *Proceedings of the Workshop on Lernen-Wissen-Adaptivität (LWA 2008)*, Würzburg, Germany, 2008, pp. 1–8. GI.
- [3] C. Fremerey, M. Müller, F. Kurth, and S. Ewert. Automatic mapping of scanned sheet music to audio recordings. In J. P. Bello, E. Chew, and D. Turnbull, eds., *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, USA, 2008, pp. 413–418. ISMIR.
- [4] F. Kurth, D. Damm, C. Fremerey, M. Müller, and M. Clausen. A framework for managing multimodal digitized music collections. In *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, Aarhus, Denmark, 2008, *LNCS 5173*, pp. 334–345. Springer.

- [5] F. Kurth and M. Müller. Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, 2008.

[6] F. Kurth, M. Müller, A. Ribbrock, T. Röder, D. Damm, and C. Fremerey. A prototypical service for real-time access to local context-based music information. In *Proc. ISMIR, Barcelona, Spain*, 2004.

30.8.3 Motion Representation and Retrieval

Investigators: Meinard Müller and Andreas Baak

In the last years, various algorithms have been proposed for automatic classification and retrieval of motion capture data. Here, one main difficulty is due to the fact that similar types of motions may exhibit significant spatial as well as temporal variations. To cope with such variations, previous algorithms often rely on warping and alignment techniques that are computationally time and cost intensive. In [1], we present a novel keyframe-based algorithm that significantly speeds up the retrieval process and drastically reduces memory requirements. In contrast to previous index-based strategies, our recursive algorithm can cope with temporal variations. In particular, the degree of admissible deformation tolerance between the queried keyframes can be controlled by an explicit stiffness parameter, see Fig. 30.29. While our algorithm works for general multimedia data, we concentrate on demonstrating the practicability of our concept by means of the motion retrieval scenario. Our experiments show that one can typically cut down the search space from several hours to a couple of minutes of motion capture data within a fraction of a second.

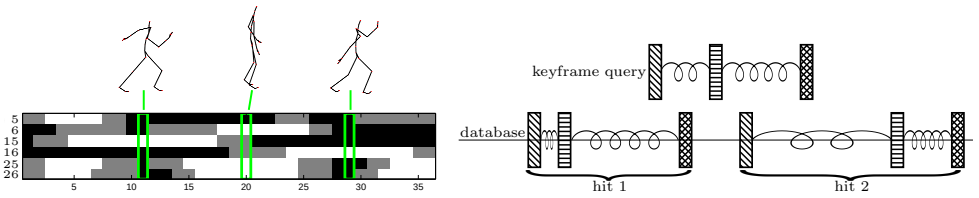


Figure 30.29: **Left:** Three characteristic key poses in a skier exercise motion along with a corresponding motion class pattern based on relational motion features. **Right:** Keyframe-based retrieval in the presence of temporal deformations.

In [3], we present a genetic learning algorithm to derive discrete patterns that can be used for classification and retrieval of 3D motion capture data. Based on boolean motion features, the idea is to learn motion class patterns in an evolutionary process with the objective to discriminate a given set of positive from a given set of negative training motions. Here, the fitness of a pattern is measured with respect to precision and recall in a retrieval scenario, where the pattern is used as a motion query. Our experiments show that motion class patterns can automate query specification without loss of retrieval quality.

Finally, in [2], we investigate how a multilinear model can be used to represent human motion data. Based on technical modes (referring to degrees of freedom and number of frames) and natural modes that typically appear in the context of a motion capture session (referring

to actor, style, and repetition), the motion data is encoded in form of a high-order tensor. This tensor is then reduced by using N-mode singular value decomposition. Our experiments show that the reduced model approximates the original motion better then previously introduced PCA-based approaches. Furthermore, we discuss how the tensor representation may be used as a valuable tool for the synthesis of new motions.

References

- [1] A. Baak, M. Müller, and H.-P. Seidel. An efficient algorithm for keyframe-based motion retrieval in the presence of temporal deformations. In M. S. Lew, A. Del Bimbo, and E. M. Bakker, eds., *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval*, Vancouver, British Columbia, Canada, 2008, pp. 451–458. ACM.
- [2] B. Krüger, J. Tautges, M. Müller, and A. Weber. Multi-mode tensor representation of motion data. *Journal of Virtual Reality and Broadcasting*, 5(5):1–13, 2008.
- [3] M. Müller, B. Demuth, and B. Rosenhahn. An evolutionary approach for learning motion class patterns. In G. Rigoll, ed., *Pattern Recognition*, Munich, Germany, 2008, *LNCS 5096*, pp. 365–374. Springer.

30.9 General Appearance Acquisition and Computational Photography

Coordinator: Hendrik P. A. Lensch

The research in our group covers the areas of computational photography, image-based modeling and rendering, 3D scanning and volume reconstruction with the goal to develop acquisition systems and algorithms for digitizing and reproducing the appearance of real-world objects, ranging from small figurines, entire streets, to planetary nebulae.

One central problem in computer graphics is the synthesis of realistic images that are indistinguishable from real photographs. While the basic theory behind rendering such images is well understood and has been turned into a large number of efficient rendering techniques, the realism of the outcome depends largely on the quality of the scene and material description. Accurate input is required for geometry, illumination and reflective properties. An efficient way to obtain realistic models is through measurement of scene attributes from real-world objects by inverse rendering. The attributes are estimated from real photographs by inverting the rendering process.

3D Geometry Reconstruction Traditional structured light 3D scanning systems are designed to capture the geometry for bright diffuse surfaces of moderate complexity. Shiny or translucent materials, e.g. metals or marble, or objects with high depth complexity typically corrupt the estimated 3D geometry producing noise or even holes in the reconstructed surface. By designing novel capturing systems, specialized illumination patterns, and appropriate reconstruction algorithms we are able to capture the precise 3D geometry even of uncooperative static as well as dynamic objects.

Appearance Measurement The research group further focuses on developing photographic techniques for measuring the scene's reflection properties. A so-called reflectance field captures the light transport within a scene such that all local and global illumination

effects, highlights, shadows, interreflections, or caustics, are recorded and can be re-rendered under arbitrary illumination. The envisioned techniques should be general enough to cope with arbitrary materials, with scenes with high depth complexity such as trees, and should allow capturing in arbitrary environments, i.e. outside a measurement laboratory.

Computational Photography A third thread of research of this group is computational photography with the goal to develop optical systems augmented by computational procedures; by jointly designing the capturing apparatus, i.e., the optical layout of active or passive devices such as cameras, projectors, beam-splitters, etc., together with the capturing algorithm and appropriate post-processing. Such combined systems are used to increase image quality, e.g., by removing image noise or camera shake, to emphasize or extract scene features such as edges or silhouettes by optical means, or to reconstruct volumetric 3D structures from images. We have developed computational photography techniques for advanced optical microscopy, large scale scene acquisition, and even astronomical imaging.

30.9.1 3D Scanning of Uncooperative Materials

Investigators: Tongbo Chen, Martin Fuchs, Matthias B. Hullin, and Hendrik P. A. Lensch

Polarization and Phase-shifting for 3D Scanning

3D scanning of translucent objects can be very difficult because of the subsurface light transport. The signal that is projected onto the object becomes weaker since it spreads out beneath the surface. Highly transmissive objects also introduce a systematic bias in the range measurement, because the peak intensity is not observed at the surface but at some point beneath it. Range measurements are furthermore polluted by interreflections and light scattering towards the measurement point.

To enable reliable range measurements on translucent objects, it is necessary to separate the first surface reflection from all the scattering effects mentioned above. One possible way to achieve this is to exploit the fact that *polarized light* becomes depolarized while undergoing multiple scattering. Therefore projecting polarized light patterns and computing the difference between images captured with a polarization filter at orthogonal orientations removes most of the subsurface scattering effects. Another method is *algorithmic descattering* which was presented by Nayar et al. in [4].

In [1, 2] we analyzed the descattering properties of polarization and phase-shifting. We show the advantages and disadvantages of both methods, and combine them to obtain better results.

Fluorescent Immersion Range Scanning

High-quality 3D scanning of optically challenging materials calls for novel acquisition techniques. While there have been many approaches addressing certain material and/or geometry constellations, none of them were able to obtain surface scans of arbitrary materials, regardless of their light transport properties. In [3] we introduce an extension of a traditional laser scanning setup that allows for robust surface scanning of objects made of translucent, transparent, glossy and very dark materials. The key idea is to observe the rays of the structured light source before they reach the object, and to disregard any light that is being reflected from the

Figure 30.30: By combining phase-shifting and polarization our method faithfully captures the 3D geometry of very translucent objects such as this alabaster Venus figurine (height ≈ 19cm).

surface. We achieve this by embedding our object in a participating medium, in particular a fluorescent dye in water, which scatters incoming light rays only once. By performing a simple thresholding on the space-time gradient, we obtain surface geometry of many different classes of materials at a high resolution and low noise figures.

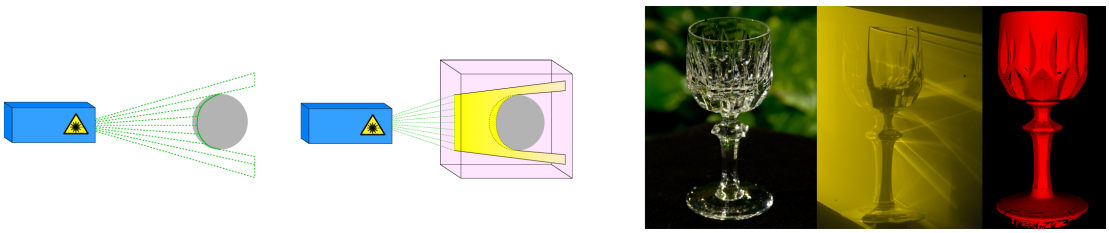


Figure 30.31: Fluorescent immersion range scanning. From left to right: Traditional laser scanner, light sheet range scanner, photo of a glass goblet, image taken with our setup, reconstructed geometry.

References

- [1] T. Chen, H. P. A. Lensch, C. Fuchs, and H.-P. Seidel. Polarization and phase-shifting for 3d scanning of translucent objects. In *2007 IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07. - Vol. 4*, Minneapolis, MN, USA, 2007, pp. 1829–1836. IEE.
- [2] T. Chen, H.-P. Seidel, and H. P. A. Lensch. Modulated phase-shifting for 3D scanning. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage, Alaska, USA, 2008. IEEE Computer Society.
- [3] M. B. Hullin, M. Fuchs, I. Ihrke, H.-P. Seidel, and H. P. A. Lensch. Fluorescent immersion range scanning. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2008)*, Los Angeles, USA, 2008, vol. 27, pp. 87:1–87:10. ACM.

[4] S. K. Nayar, G. Krishnan, M. D. Grossberg, and R. Raskar. Fast separation of direct and global components of a scene using high frequency illumination. *ACM TOG (Proc. ACM SIGGRAPH 2006)*, 25(3):935–944, 2006.

30.9.2 Acquisition and Rendering of Scene and Material Appearance

Investigators: Martin Fuchs, Christian Fuchs, Volker Blanz, and Hendrik P. A. Lensch

The approaches that can be used to accurately measure the appearance of a real-world scene can be classified along the required a-priori assumptions on the scene geometry and reflectance. In recent years, image-space methods that try to establish as many of these as possible from direct observation in digital pictures have shown considerable success and general applicability. Most direct in that respect is the concept of the *reflectance field* [1] – it models the light transport as the relation between the radiance that is cast into the scene on a given incoming light ray and the radiance which can be observed on an outgoing ray, both chosen from a 4D parameterization. This amounts to an 8D data structure.

The main issue with reflectance fields remains the high amount of data that needs to be recorded for a faithful representation. Even after reducing the problem to image-space relighting (external illumination is assumed distant, the camera remains fixed for all renderings, thus a 4D data structure is sufficient), tens of thousands of input images need to be acquired in brute-force approaches [5] for faithful reconstruction. We have therefore developed methods that reduce the required measurement effort significantly.

Adaptive Sampling of Reflectance Fields

a) 256 input pictures, point sampling. Note speckled appearance of highlights.

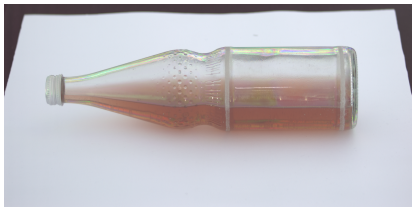
b) 256 input pictures, pre-filtered

c) Setup used. Spotlight projectors with variable aperture diameter illuminate the scene indirectly.

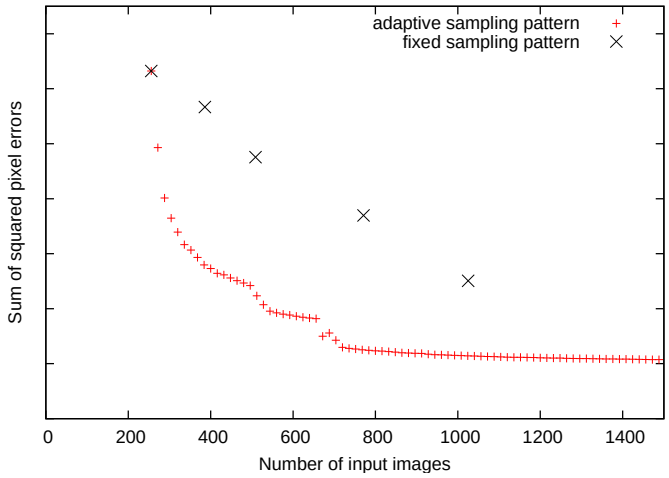
Figure 30.32: Sampling reflectance fields with appropriate pre-filtering (b) mitigates artifacts introduced by aliasing (a) even in cases where the reflectance field was only sparsely sampled. An appropriate pre-filtering kernel can be created in an indirectly illuminated, dark tent (c).

In setups with conventional point light sources, sparse reflectance field recordings ($n \approx 200 \ldots 300$) expose visible deficits in the reconstruction quality (see Figure 30.32 (a)). In the context of sampling theory, these can be understood as aliasing artifacts. Interpreting the incident illumination during capture as sampling kernels enables us to choose extended, smooth illuminations that cause appropriate pre-filtering to happen while the reflectance field is acquired in a way that effectively suppresses aliasing artifacts (see Figure 30.32 (b)). We can generate such illumination kernels in a simple setup making use of indirect illumination (Figure 30.32 (c)).

a) Rendering of a regularly sampled reflectance field with 386 input pictures.



b) Rendering of an adaptively sampled reflectance field with 384 input pictures and appropriately chosen pre-filter kernels.



c) Comparison of adaptive sampling with regular sampling (constant pre-filter size).

Figure 30.33: Benefits of adaptive sampling. Due to undersampling, the rendering in dataset (a) exposes a speckled appearance in the highlight regions. With appropriately chosen pre-filter kernels and sampling positions (b), they disappear, but the colorful highlights at the top of the bottle remain sharp and crisp. As (c) shows, the error compared to a ground-truth measurement with 10 000 input images decreases dramatically faster with adaptive sampling than with regular, fixed pattern sampling.

While extended illuminations effectively inhibit aliasing, they achieve that by suppressing high frequencies, which blurs out shadow edges and highlights in rendered result images, so, for some scenes, dense sampling can still be a requirement. However, this is not equally true for all regions in the incident illumination domain. We therefore propose an adaptive sampling scheme, which during recording establishes estimates on the required sampling density in different regions for a given scene. By distributing the samples accordingly, we can massively reduce the required measurement effort (see Figure 30.33).

We have published both contributions, appropriate pre-filtering and our adaptive sampling

scheme, in the ACM Transactions on Graphics [2].

Reflectance Field Super-resolution

a) 230 input pictures, linear interpolation

b) 230 input pictures, upsampled to 3 547

c) 14 116 input pictures (ground truth)

Figure 30.34: Effect of super resolution processing for reflectance fields. A plain linear interpolation (left) leaves gaps in highlights and fails to reproduce soft shadows. The improved interpolation creates a high quality (center), which comes close to a ground truth reference (right).

For some cases, e.g. a mirror ball, neither pre-filtering nor adaptive reflectance field sampling is effective due to the high-frequency response. We have therefore developed an upsampling technique for reflectance fields [3], presented at Eurographics 2007. It works by analyzing a sparsely sampled field, and decomposing a reflectance field into layers of different frequency characteristics: low-frequency effects, such as near-diffuse reflectance components, can be linearly interpolated. High-frequency effects, though, require an explicit treatment: highlights positions are established and warped along the surface geometry after performing optical flow contributions. Shadowed regions are estimated separately, and after performing a level-set blend in order to obtain an in-between image, used to determine whether a given pixel at output resolution ought to be shadowed or not. By combining a local extrapolation of the apparent pixel BRDF with texture priors on the appearance of shadow boundaries, accurate reconstructions become possible (see Figure 30.34).

A Passive Reflectance Field Display

Reflectance fields can be used to preserve the appearance of a scene in digital form, and to relight the scene synthetically under arbitrary illumination. Capturing the incident illumination at a specific point, one can predict the appearance of the object at that position [7].

Investigating the rendering equation, we have found that its main mathematical operators, integration and multiplication, can be implemented by simple optical components. So, instead of having to measure the room illumination, rendering a picture and displaying it on an electronic display, we can directly transform the incident illumination into the desired output picture (see Figure 30.35). This creates a display that in real-time adapts its picture to its surroundings, does not consume electricity and can be reconfigured to show different scenes by exchanging a single transparency slide in its inside – a transparency slide that is printed with a standard, off-the shelf ink jet printer.

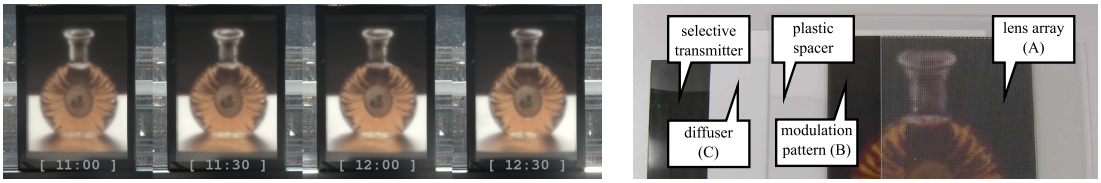


Figure 30.35: A 4D Passive Reflectance Field Display: mounted to a window in daylight (left), it renders a scene in the outside illumination. As time progresses, and the sun traverses the sky, highlights and shadows move naturally in the generated image. The embedded construction consists of few off-the-shelf components (right).

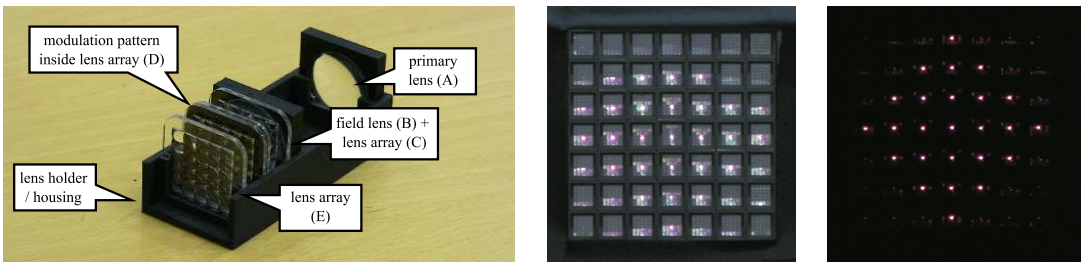


Figure 30.36: The construction of a single pixel of a 6D reflectance field display requires custom-manufactured components. Single pixels (left) are stacked together to form a display (center) which can display simple patterns (right) with viewer and illumination dependency.

Our illumination-dependent display construction was inspired by integral photography [6], which uses a system of optical lenses to create a view-dependent appearance. By augmenting the previous construction with additional components (see Figure 30.36), we can also create view-dependent effects while maintaining the illumination-dependence, albeit for simpler, abstract geometrical patterns as output images. Thus, we can passively visualize a data structure of six dimensions – two for the image extent, two for the viewer position (angle to the display), and two for the direction of the incident illumination.

This project is joint work with Ramesh Raskar (now at MIT Media Lab) and started during an internship of Martin Fuchs at Mitsubishi Electric Research Laboratories (MERL), Cambridge, USA, in summer 2007. The work has been published at ACM SIGGRAPH 2008 [4].

Relighting Objects from Image Collections

Measuring the spatially varying BRDFs (reflectance properties) of objects is usually done in an environment where the lighting can be precisely controlled. If done in uncontrollable lighting (e.g. outside) the lighting conditions typically must be known, for example by measuring them

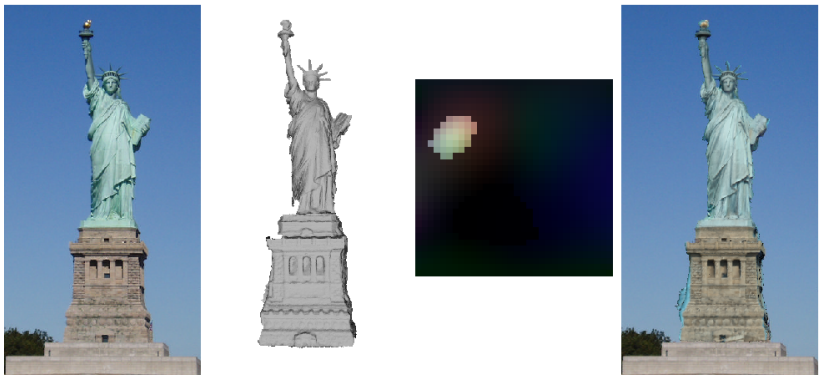


Figure 30.37: Overview of our reconstruction pipeline. From left to right: an example image taken from Flickr, estimated geometry, recovered environment map for this particular input view, and rendered result using the geometry, environment and estimated reflectance properties.

with a light probe. Furthermore almost all previous approaches need the object's geometry as input.

Our system [8] is the first to recover object geometry, spatially varying reflectance properties, and the illumination conditions for each input image from just a small number of photographs. Since the input images may be taken under different (uncontrolled) illumination and from different viewpoints our system is even able to use images we obtained from a community photo collection like Flickr.

References

[1] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, 2000, pp. 145–156. ACM Press/Addison-Wesley Publishing Co.

- [2] M. Fuchs, V. Blanz, H. P. A. Lensch, and H.-P. Seidel. Adaptive sampling of reflectance fields. *ACM Transactions on Graphics (TOG)*, 26(2):1–18, 2007.
- [3] M. Fuchs, H. P. A. Lensch, V. Blanz, and H.-P. Seidel. Superresolution reflectance fields: Synthesizing images for intermediate light directions. In D. Cohen-Or and P. Slavík, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Prague, Czech Republic, 2007, vol. 26(3), pp. 447–456. Blackwell.
- [4] M. Fuchs, R. Raskar, H.-P. Seidel, and H. P. A. Lensch. Towards passive 6D reflectance field displays. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 27(3):58, 2008.

[5] T. Hawkins, P. Einarsson, and P. Debevec. A dual light stage. In *Rendering Techniques 2005 (Proc. Eurographics Symposium on Rendering)*, 2005, pp. 91–98.

[6] G. Lippmann. Epreuves reversibles donnant la sensation du relief. *Journal of Physics*, 7:821–825, 1908.

[7] S. K. Nayar, P. N. Belhumeur, and T. E. Boult. Lighting sensitive display. *ACM Transactions on Graphics*, 23(4):963–979, 2004.

Figure 30.38: Background estimation. Using graph cuts, the background of a set of photographs (right) is obtained from a minimum cost composite (center) of the given input photographs (left).

- [8] H.-P. Seidel. Relighting objects from image collections. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami Beach, Florida, USA, 2009, p. 8. IEEE Computer Society.

30.9.3 Image Processing

Investigators: Boris Ajdin, Miguel Granados, Matthias B. Hullin, Christian Fuchs, and Hendrik P. A. Lensch

Background Estimation / HDR Deghosting

In our paper [2], we address the problem of reconstructing the background of a crowded scene from a set of photographs, which were exposed using the same camera configuration and illumination conditions. An example of such an input and the estimated background is presented in Figure 30.38.

Our approach is to define the background image as a composite of the input photographs. Each possible composite is assigned a cost, and the resulting cost function is minimized using graph cuts. We penalize deviations from the following two model assumptions: background objects are stationary, and background objects are more likely to appear across the photographs. In order to avoid background estimations that cut through objects, we constraint the solution space to only allow transitions which are locally consistent with the input photographs. Lastly, by first transforming the input photographs to the radiance domain using the camera's transfer function, the same approach can be used to remove ghosting artifacts during the reconstruction of high dynamic range images from multi-exposure sequences.

Demosaicing

Most cameras present on the market nowadays consist of various lens elements placed in front of a CCD or CMOS sensor. Each pixel of the sensor is able to convert incoming radiance into digital value, which approximates the number of photons falling onto the sensor surface. However, typical sensing elements are unable to distinguish between different light wavelengths – only the radiance integrated over all wavelengths is measured. In oder to

separate wavelength bands, an additional color filter is added in front of each pixel which lets only a certain range of wavelengths through. To mimic the human visual system, red, green and blue color filters are the most common ones used. As a final result of this process a color mosaic image is obtained, where each pixel has only one color channel captured (either red, green or blue). In order to produce a full RGB image we need to employ a so-called demosaicing algorithm, which interpolates the missing colors. In our paper [1], we investigate the properties of previous approaches to demosaicing and introduce a new method, which makes use of a pixel neighborhood to determine the missing data. Unlike other authors, in our work we explore the error metric in a global sense within the pixel neighborhood, trying to interpolate in the best possible way. We present one result obtained by our method in Figure 30.39.

Figure 30.39: Comparison images for previous "state-of-the-art" method (left), and the proposed method (right).

References

- [1] B. Ajdin, M. B. Hullin, C. Fuchs, H.-P. Seidel, and H. P. A. Lensch. Demosaicing by smoothing along 1d features. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.
- [2] M. A. Granados V., H.-P. Seidel, and H. P. A. Lensch. Graphics interface. In C. Shaw and L. Bartram, eds., *Proceedings of the Graphics Interface 2008 Conference*, Windsor, Ontario, Canada, 2008, ACM International Conference Proceeding Series, pp. 33–40. ACM Press.

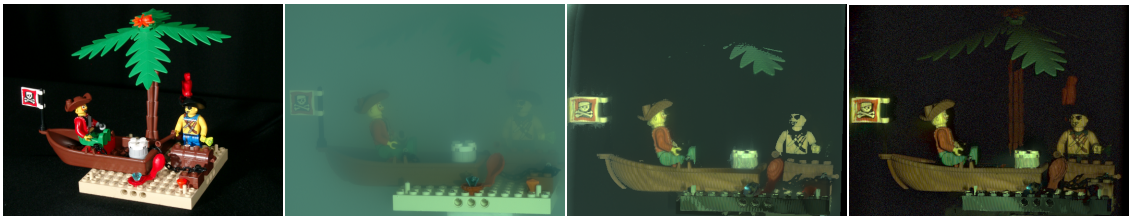
30.9.4 Light Transport in Volumes

Investigators: Christian Fuchs, Martin Fuchs, Andrei Linţu, Matthias B. Hullin, and Hendrik P. A. Lensch

Combining Descattering and Confocal Imaging

In section 30.9.1 we developed methods for eliminating the subsurface light transport in translucent objects (so-called *descattering*) with the goal of enabling 3D scanning of this

class of objects. In [1] we apply similar techniques to facilitate acquisition of an object which is embedded in a translucent medium. The classic application for this method is looking through fog or murky water.



(a) (b) (c) (d)

Figure 30.40: Example results from our descattering approach: *(a)* shows a photograph of the object that is immersed in a fishtank filled with dilute milk. After that, an image taken under normal flood-light illumination is shown in *(b)*. Images (c) and (d) show the advantage of combining confocal imaging and descattering. With just confocal imaging *(c)* some details can not be recovered. Combining both methods *(d)* leads to a sharper image with more detail and better color reproduction.

We analyze two widespread techniques, *algorithmic descattering* [6] and synthetic aperture *confocal imaging* [4]. We combine both methods with the goal of computing cross-sectional images in translucent media with higher contrast and better resolution than previously possible.

Direct Volumetric Acquisition of Homogeneous Refractive Objects

As described in section 30.9.1, embedding a traditional 3D scanning setup in a fluorescent medium allows for acquiring the geometry of objects made from a wide range of materials. In the case of homogeneous transparent materials such as clear glass, we can even go a step further and acquire entire slices through their geometry at a time. This can be accomplished by matching the refractive index of the material with the one of the fluorescent medium. The result is a binary density volume, similar to what one would obtain e. g. from computerized tomography, but without the need for tomographic reconstruction. This technique was also published in [3].

Direct Visualization of Real-world Light Transport

In [2], we demonstrate that by observing the propagation of light inside a volume, we can obtain valuable insights about the function of optical elements and the reflectance of surfaces. For instance, it is possible to intuitively inspect the reflection lobes of different materials.

3D Reconstruction of Reflection Nebulae from a Single Image

The light transport on a entirely different scale has been investigated in [5]. Here, we analyzed how the 3D structure of planetary nebulae can be reconstructed. The main problem in

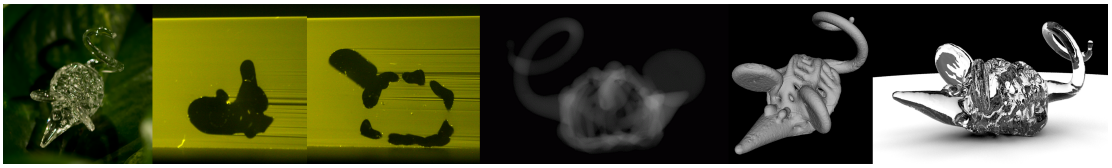


Figure 30.41: Direct volume scanning of homogeneous clear objects. From left to right: photo of glass mouse, two example slices as captured by our system, volume rendering, extracted surface mesh, glass shader rendering.

here is that observations from earth always show the planetary object from the same view. Traditional volumetric reconstruction methods such as tomography can therefore not be applied. We addressed the reconstruction problem through analysis by synthesis, trying to optimize the volume density by comparing a rendering of the current volume and comparing it to the observations. A couple of additional information is used to stabilize the estimation process: the use of multi-wavelength observations to separate the effects of reflections, emission and absorption, and regularization in the form of assume rotational symmetry or smoothness perpendicular to the observation direction.

References

- [1] C. Fuchs, M. Heinz, M. Levoy, H.-P. Seidel, and H. P. A. Lensch. Combining confocal imaging and descattering. *Computer Graphics Forum (Proc. Eurographics Symposium on Rendering)*, 27(4):1245–1253, 2008.

- [2] M. B. Hullin, M. Fuchs, I. Ihrke, B. Ajdin, H.-P. Seidel, and H. P. A. Lensch. Direct visualization of real-world light transport. In O. Deussen, D. Keim, and D. Saupe, eds., *13th International Fall Workshop on Vision, Modeling and Visualization*, Konstanz, Germany, 2008, pp. 363–371. Akademische Verlagsgesellschaft AKA.

- [3] M. B. Hullin, M. Fuchs, I. Ihrke, H.-P. Seidel, and H. P. A. Lensch. Fluorescent immersion range scanning. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2008)*, Los Angeles, USA, 2008, vol. 27, pp. 87:1–87:10. ACM.

[4] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. Bolas. Synthetic aperture confocal imaging. *ACM TOG (Proc. ACM SIGGRAPH 2004)*, 23(3):825–834, 2004.

- [5] A. Lintu, L. Hoffmann, M. Magnor, H. P. A. Lensch, and H.-P. Seidel. 3d reconstruction of reflection nebulae from a single image. In H. P. A. Lensch, B. Rosenhahn, H.-P. Seidel, P. Slusallek, and J. Weickert, eds., *Vision, Modeling, and Visualization*, Saarbrücken, Germany, 2007, pp. 109–116. MPI Informatik, AKA.

[6] S. K. Nayar, G. Krishnan, M. D. Grossberg, and R. Raskar. Fast separation of direct and global components of a scene using high frequency illumination. *ACM TOG (Proc. ACM SIGGRAPH 2006)*, 25(3):935–944, 2006.

30.10 Advanced Global Illumination and Realtime Realistic Image Synthesis

Coordinators: Elmar Eisemann, Thorsten Grosch, and Karol Myszkowski

The goal of this research area is the generation of photorealistic images and animations with a special focus on *Global Illumination*. Using Global Illumination significantly increases the realism of computer generated scenes: When displaying 3D models without natural illumination, they often look synthetic — a realistic look can only be achieved with correct illumination and shadows.

Therefore, several methods were developed to improve the state-of-the-art Global Illumination algorithms: One example is the *Photon Mapping* method, where light is simulated by tracing small particles (photons). The stochastic nature of this process incurs several visual artifacts, like noise and bias problems at geometric edges. To improve the display quality, an extension of the Photon Mapping algorithm was developed, which is described in Section 30.10.1. Instead of rendering only single images, generating animations is more interesting. For such a rendered video, lossy compression methods like MPEG are often used to reduce the amount of storage. Section 30.10.2 describes a rendering method that is *adapted to MPEG* and avoids the computation of unnecessary details that would be removed by the MPEG compression afterwards.

In contrast to precomputing images or videos, another research direction is the computation of Global Illumination in *real-time*. Section 30.10.3 explains how the current graphics hardware (GPU) can be used to compute time-consuming global illumination effects like indirect light and indirect shadows at interactive to real-time framerates. The main idea is to use approximations for indirect visibility (*Imperfect Shadow Maps*) which are almost not perceivable for the human visual system. The focus in this project is the computation of Global Illumination in completely *dynamic* scenes, for example in computer games. A special problem in this area is the real-time computation of soft shadows. Here, a new method was developed, so called *Convolution Soft Shadow Maps*, which allows the display of a soft shadow at several hundred frames per second. This method is described in Section 30.10.4.

Up to this point, all the presented lighting computations were performed in vacuum. If a *participating medium* is considered, all lighting computations become more complicated. To efficiently solve the light transport in such volumetric objects, the so-called *Eikonal Rendering* was developed. This allows the simulation of complicated light refractions at real-time framerates on the GPU, as shown in Section 30.10.5.

Besides Global Illumination, several other projects are related to the generation of realistic images. The GPU is also used in a project for real-time rendering of height fields, as described in Section 30.10.6. Voxel-based representations are viable alternatives for scene modeling with respect to commonly used mesh. In Section 30.10.7 we address the problem of efficient GPU rendering of such scene representations, while in Section 30.10.8 we present a GPU-based algorithm for isosurface extraction from volumetric data. Finally, Section 30.10.9 explains a method that generates realistic animations of floating water and smoke from a single image by transferring the information from a simple example animation to the still image.

30.10.1 Global Illumination using Photon Ray Splatting and Radiance Caching

Investigators: Robert Herzog and Karol Myszkowski

Computing the global illumination in synthetic scenes is one of the most appealing but also most costly components in realistic image synthesis. One efficient way of computing the global illumination is photon density estimation. Photon density estimation is capable of handling all kinds of light transports. However, the method introduces various types of bias as well as low-frequency noise. One type of bias is the geometric bias which is due to the assumption that the domain for density estimation is unbounded and planar.

Our new metric [1, 2] performs photon density estimation in ray space independent of the underlying geometry and eliminates the geometric bias. Besides, our method also reduces variance in the density estimate. In order to be still efficient we propose a novel data structure for searching the nearest neighbors points for a conical frustum and a radiance cache for sparse sampling and filtering in image space.

Our method proposed in [1] still suffers from proximity bias which is dominated by the high-frequency visibility term. In [4] we approximate the mutual visibility inside a photon-ray density-estimation footprint by traversing a 3D voxelization of the synthetic scene for a cylindrically-shaped footprint. The voxelization is efficiently done on the GPU by means of 3D rasterizaton with an orthographic camera. We achieve high-quality results for diffuse light transport even for direct illumination using only a few thousand photons. One example is shown in Figure 30.42.

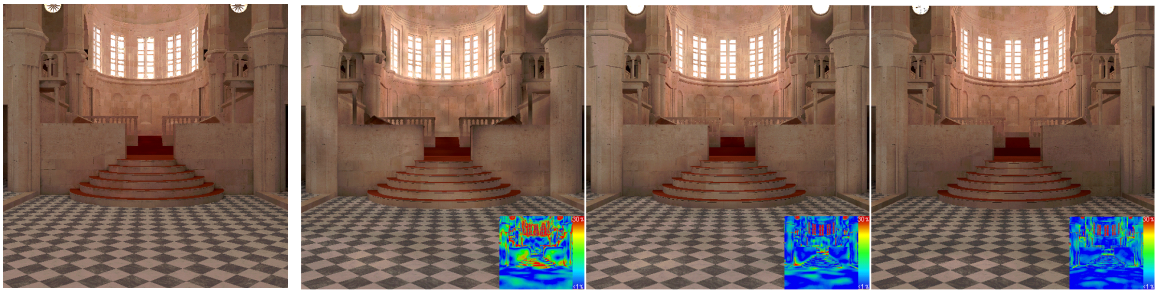


Figure 30.42: From left to right: ground-truth, traditional photon density estimation, our photon ray splatting [1], ray splatting with caching and adaptive noise filtering.

Radiance caching on 3D object surfaces in combination with photon mapping is the state-of-the-art method for computing high-quality production-style images and movies in the film industry and for lighting design. Nevertheless, computing high-quality results with radiance caching is still far from being interactive and besides requires an experienced user to set various parameters. Our method [3] is built on current state-of-the-art radiance caching and improves the method in several aspects: First, we propose a combination of the radiance cache with the lightcuts algorithm, which automatically adapts to lighting complexity. This way, we do not only reduce the computation time but also reduce the space of user parameters. Second, we replace the traditional cache gathering by an anisotropic cache splatting approach, which

provides better spatial coherence and better search efficiency in particular for high-resolution images.

References

- [1] R. Herzog, V. Havran, S. Kinuwaki, K. Myszkowski, and H.-P. Seidel. Global illumination using photon ray splatting. In D. Cohen-Or and P. Slavik, eds., *The European Association for Computer Graphics 28th Annual Conference: EUROGRAPHICS 2007*, Prague, Czech Republic, 2007, *Computer Graphics Forum*, vol. 26(3), pp. 503–513. Blackwell.
- [2] R. Herzog, V. Havran, S. Kinuwaki, K. Myszkowski, and H.-P. Seidel. Global illumination using photon ray splatting. Research Report MPI-I-2007-4-007, MPI Informatik, Saarbruecken, Germany, 2007.
- [3] R. Herzog, K. Myszkowski, and H.-P. Seidel. Anisotropic radiance-cache splatting for efficiently computing high-quality global illumination with lightcuts. In M. Stamminger and P. Dutré, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, München, Germany, 2009, vol. 28, pp. 259–268. Wiley-Blackwell.
- [4] R. Herzog and H.-P. Seidel. Lighting details preserving photon density estimation. In M. Alexa, T. Ju, and S. Gortler, eds., *The 15th Pacific Conference on Computer Graphics and Application*, Maui, Hawaii, USA, 2007, pp. 407–410. IEEE Computer Society.

30.10.2 Integrating Rendering and Video Compression

Investigators: Robert Herzog and Karol Myszkowski

Currently, 3D animation rendering and video compression are completely independent processes, even if rendered frames are streamed on-the-fly within a client-server platform. In such a scenario, which may involve time-varying transmission bandwidths and different display characteristics at the client side, dynamic adjustment of the rendering quality to such requirements can lead to a better use of server resources. In [1], we present a framework where the renderer and MPEG codec are coupled through a straightforward interface that provides precise motion vectors from the rendering side to the codec and perceptual error thresholds for each pixel in the opposite direction. The perceptual error thresholds take into account bandwidth-dependent quantization errors resulting from the lossy compression as well as image content-dependent luminance and spatial contrast masking. The availability of the discrete cosine transform (DCT) coefficients at the codec side enables to use advanced models of the human visual system (HVS) in the perceptual error threshold derivation without incurring any significant cost. Those error thresholds are then used to control the rendering quality and make it well aligned with the compressed stream quality (for an example see Figure 30.43). In our prototype system we use the lightcuts technique developed by Walter et al., which we enhance to handle dynamic image sequences, and an MPEG-2 implementation. Our results clearly demonstrate many advantages of coupling the rendering with video compression in terms of faster rendering.

References

- [1] R. Herzog, S. Kinuwaki, K. Myszkowski, and H.-P. Seidel. Render2MPEG: a perception-based framework towards integrating rendering and video compression. In R. Scopigno and E. Gröller,

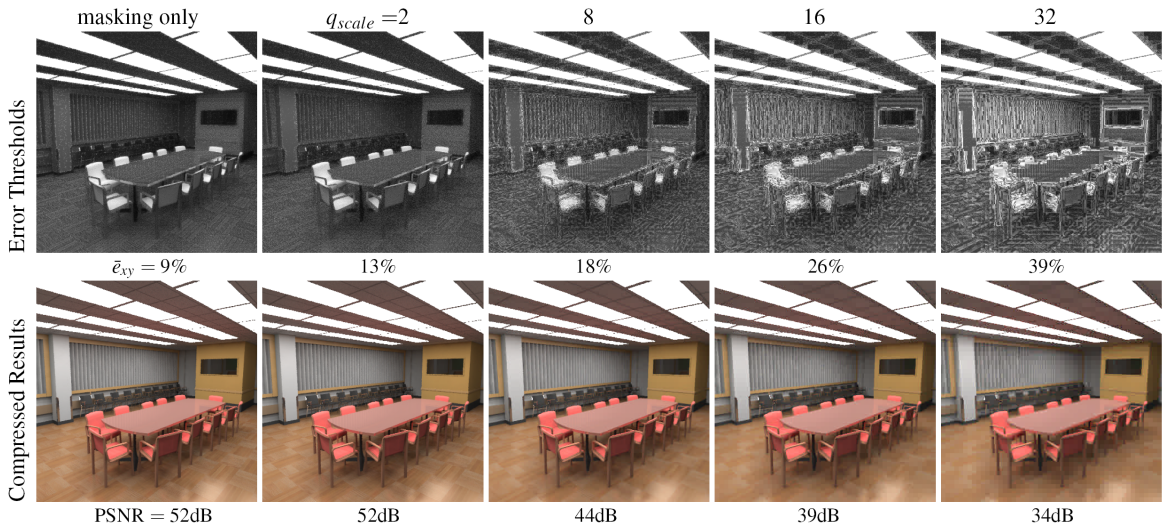


Figure 30.43: From left to right: decreasing MPEG bandwidth (q_{scale}), top: quantization depending error thresholds, bottom: MPEG-compressed results, rendered using the thresholds above, are visually indistinguishable to the reference after compression.

eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Crete, Greece, 2008, vol. 27(2), pp. 183–192. Blackwell.

30.10.3 Efficient Computation of Dynamic Indirect Illumination

Investigators: Tobias Ritschel and Thorsten Grosch

Rendering physically plausible illumination at real-time frame-rates is often achieved using approximations. We have developed two novel approaches to approximate costly indirect visibility computations as required for global illumination, giving physically plausible and visually pleasing results. Indirect visibility is needed to faithfully reproduce indirect shadows cast by indirect light. Such shadows are especially challenging to compute in dynamic (i.e. deforming) scenes or scenes with fine geometric details.

The first method [3] exploits, how indirect illumination mostly consists of smooth gradations, which tend to mask errors due to incorrect visibility. We exploit this observation using *Imperfect Shadow Maps* — low-resolution shadow maps rendered from a crude point-based representation of the scene. These are used in conjunction with a global illumination algorithm based on virtual point lights enabling indirect illumination of dynamic scenes at real-time frame rates. We demonstrate that Imperfect Shadow Maps are a valid approximation to visibility, which makes the simulation of global illumination an order of magnitude faster than using accurate visibility. Figure 30.44 shows examples rendered with our method.

The second approach [4] is based on Ambient Occlusion, for which very simple and efficient implementations are used extensively in production rendering. Ambient Occlusion methods

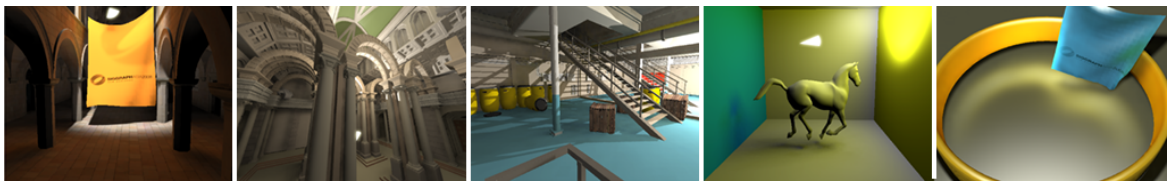


Figure 30.44: Imperfect Shadow Maps [3] render diffuse Global Illumination in dynamic scenes at interactive rates (5–18 fps).

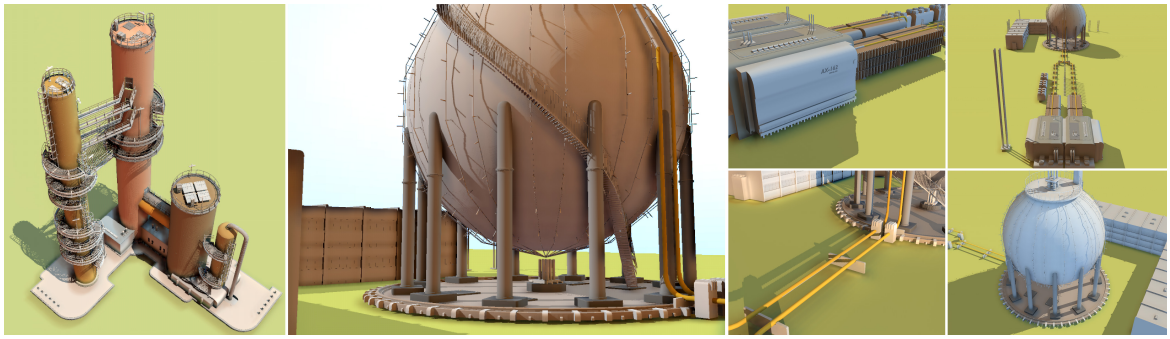


Figure 30.45: Screen Space Directional Occlusion [4] renders bouncing light and directional shadows between fine surface details (20 fps).

often precompute average visibility values and multiply them with the unoccluded illumination, which is computed in real-time with the graphics hardware. Recent methods approximate Ambient Occlusion between nearby geometry in screen space (SSAO), for example by using the information stored in the depth buffer. Our observation is, that screen-space occlusion methods can be used to compute many more types of effects than just occlusion: We compute so-called *Directional Occlusion*, where shadows with correct color and orientation can be displayed. In contrast to this, Ambient Occlusion can only display static and grey contact shadows. Additionally, approximate indirect color bleeding can be computed in image space with only a small overhead. Some example renderings are shown in Figure 30.45. Since our method works in image space, it can be combined with any other method that simulates light transport for macro structures.

In case of static scenes, visibility information can be precomputed and stored in compressed form as Coherent Shadow Maps. This allows fast visibility queries for direct light. We extended this method to so-called *Coherent Surface Shadow Maps* [2], where additional visibility information stored for every point on the surface of the scene. In this way, fast visibility tests for indirect light can be performed.

The fast, GPU-based simulation of light can also be used to *augment live images* from

a video camera with virtual objects with correct illumination and shadows. Here, a High-Dynamic-Range video camera permanently captures the incoming light, which is then used for the illumination of the virtual objects. Due to the correct illumination, the virtual objects can be seamlessly integrated into the camera image, as shown in [1]. The main focus here was the correct description of the *near field* of the illumination, allowing interactive movement of virtual objects in every region of the camera image.

References

- [1] T. Grosch, T. Eble, and S. Mueller. Consistent interactive augmentation of live camera images with correct near-field illumination. In S. N. Spencer, ed., *VRST 2007 (ACM Symposium on Virtual Reality Software and Technology)*, Newport Beach, California, USA, 2007, pp. 125–132. ACM.
- [2] T. Ritschel, T. Grosch, J. Kautz, and H.-P. Seidel. Interactive global illumination based on coherent surface shadow maps. In C. Shaw and L. Bartram, eds., *Proceedings Graphics Interface, May 2008*, Windsor, Ontario, Canada, 2008, ACM International Conference Proceeding Series, pp. 185–192. ACM.
- [3] T. Ritschel, T. Grosch, M. H. Kim, H.-P. Seidel, C. Dachsbacher, and J. Kautz. Imperfect shadow maps for efficient computation of indirect illumination. In *Proceedings of ACM SIGGRAPH Asia 2008*, Singapore, 2008. ACM.
- [4] T. Ritschel, T. Grosch, and H.-P. Seidel. Approximating dynamic global illumination in image space. In *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, Boston, MA, USA, 2009, pp. 75–82. ACM.

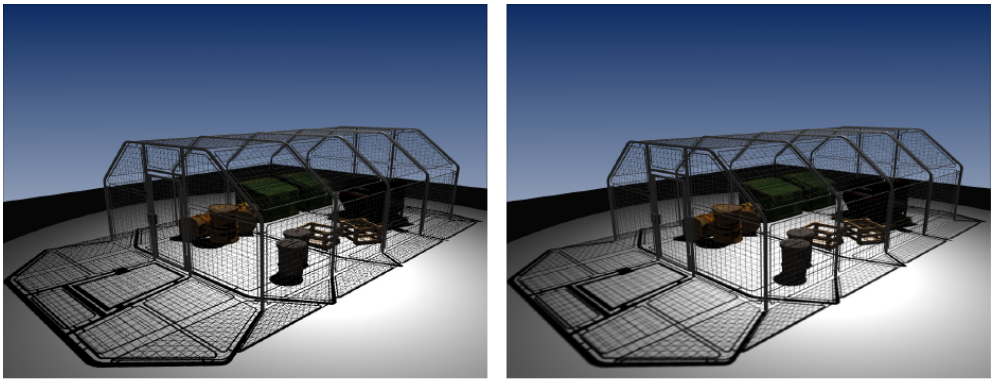
30.10.4 High-quality Shadow Map Filtering and its Applications

Investigators: Thomas Annen and Zhao Dong

Anti-aliasing for shadow mapping is a classical problem for graphics rendering research. We have developed several shadow map filtering methods which can render hard and soft shadows with both high quality and real-time performance.

The first shadow map filtering method is the *Convolution Shadow Map* (CSM)[2], a novel shadow representation that affords efficient, arbitrary linear filtering of shadows. Traditional shadow mapping is inherently non-linear w.r.t. the stored depth values, due to the binary shadow test. We linearize the problem by approximating the shadow test as a weighted summation of basis terms. We demonstrate the usefulness of this representation, and show that hardware-accelerated anti-aliasing techniques, such as tri-linear filtering, can be applied naturally to Convolution Shadow Maps. Our approach can be implemented very efficiently in current generation graphics hardware, and offers real-time frame rates. One example is shown in Figure 30.46.

The second approach is the *Exponential Shadow Map* (ESM)[3], which is a simple approach to shadow map filtering, by approximating the shadow test using an exponential function. This enables us to pre-filter the shadow map, which in turn allows for high quality hardware-accelerated filtering. Compared to previous filtering techniques, our technique is faster, consumes less memory and produces less artifacts. Some comparisons are shown in Figure 30.47.



Percentage Closer Filtering | CSM with 7x7 blur and mip-mapping

Figure 30.46: Standard percentage closer filtering does not support tri-linear filtering and suffers from severe aliasing artifacts during minification. In contrast, Convolution Shadow Maps (CSM) enable tri-linear filtering of shadows and thereby achieve effective screen-space anti-aliasing. Additional convolution can hide shadow map discretization artifacts.)

In order to achieve real-time, photo-realistic rendering of computer-generated scenes, we introduce the *Convolution Soft Shadow Map*[1] method. It is a very fast method for rendering plausible soft shadows based on convolution theory. It requires only a constant-time memory lookup, thereby enabling us to render soft shadows at hundreds of frames per second for a single area source. Environment-lit scenes can be rendered from a collection of approximating area light sources. Even though shadows are only approximate, the results are virtually indistinguishable from reference renderings, but are produced at real-time frame rates. Figure 30.48 shows some results.

References

- [1] T. Annen, Z. Dong, T. Mertens, P. Bekaert, H.-P. Seidel, and J. Kautz. Real-time, all-frequency shadows in dynamic scenes. In G. Turk, ed., *Proceedings of ACM SIGGRAPH 2008*, Los Angeles, USA, 2008, *ACM Transactions on Graphics*, vol. 27, pp. 1–8. ACM.

- [2] T. Annen, T. Mertens, P. Bekaert, H.-P. Seidel, and J. Kautz. Convolution shadow maps. In J. Kautz and S. Pattanaik, eds., *Rendering Techniques 2007: Eurographics Symposium on Rendering*, Grenoble, France, 2007, *Eurographics / ACM SIGGRAPH Symposium Proceedings*, vol. 18, pp. 51–60. Eurographics.

- [3] T. Annen, T. Mertens, H.-P. Seidel, E. Flerackers, and J. Kautz. Exponential shadow maps. In C. Shaw and L. Bartram, eds., *Proceedings of Graphics Interface 2008*, Windsor, Ontario, Canada, 2008, *ACM International Conference Proceeding Series*, vol. 34, pp. 155–161. A K Peters.

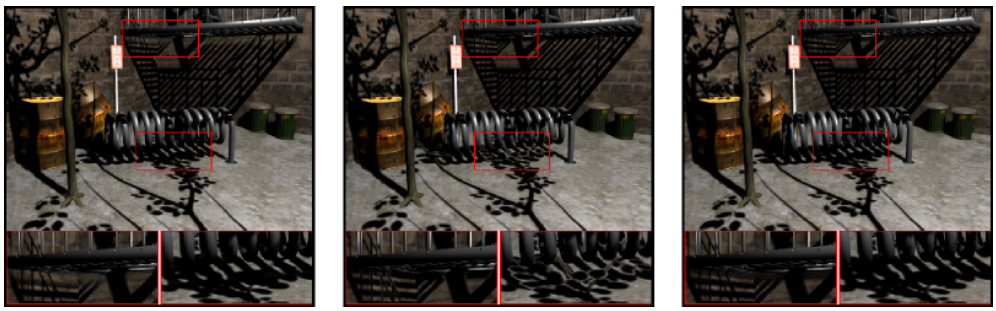


Figure 30.47: A backyard scene rendered with a $2k \times 2k$ shadow map and 5×5 Gauss filtering using (from left to right, statistics include mip-map memory): CSMs (22 FPS, 170 MB), VSMs (84 FPS, 42 MB), and ESMs (94 FPS, 21 MB). ESMs require $8\times$ less memory than CSMs and have less light leaking at contact points while their filtering quality is almost indistinguishable. Like CSMs, ESMs also avoid the high frequency light leaking artifacts seen with VSMs.

30.10.5 Efficient Light Transport in Refractive Objects

Investigators: Ivo Ihrke, Gernot Ziegler, and Art Tevs

We developed a new method for real-time rendering of sophisticated lighting effects in and around refractive objects [1]. It enables us to realistically display refractive objects with complex material properties, such as arbitrarily varying refraction index, inhomogeneous attenuation, as well as spatially-varying anisotropic scattering and reflectance properties. User-controlled changes of lighting positions only require a few seconds of update time. Our method is based on a set of ordinary differential equations derived from the eikonal equation, the main postulate of geometric optics. This set of equations allows for fast casting of bent light rays with the complexity of a particle tracer. Based on this concept, we also propose an efficient light propagation technique using adaptive wavefront tracing. Efficient GPU

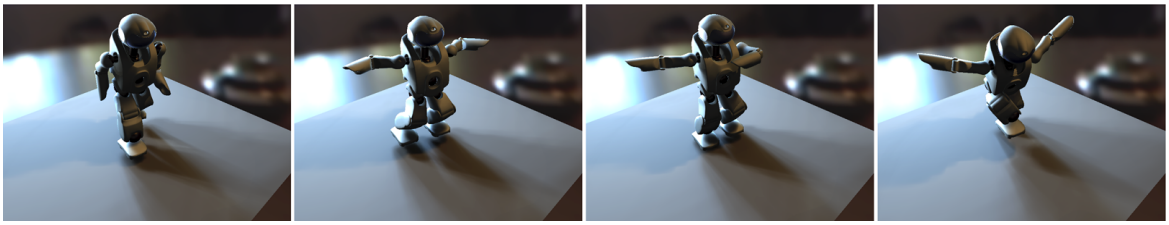


Figure 30.48: Based on the Convolution Soft Shadow Map method, a fully dynamic animation of a dancing robot under environment map lighting is rendered at 29.4 fps without any precomputation. Incident radiance is approximated by 30 area light sources (256×256 shadow map resolution each).

Figure 30.49: Real-time renderings of complex refractive objects. (left) glass with red wine casting a colorful caustic, 24.8 fps. (middle) Amberlike bunny with black embeddings showing anisotropic scattering and volume caustics in the surrounding smoke and its interior, 13.0 fps. (right) Rounded cube composed of three differently colored and differently refracting kinds of glass showing scattering effects and caustics in its interior, 6.4 fps.

implementations for our algorithmic concepts enable us to render visual effects that were previously not reproducible in this combination in real-time. See results on Figure 30.49.

References

- [1] I. Ihrke, G. Ziegler, A. Tevs, C. Theobalt, M. Magnor, and H.-P. Seidel. Eikonal rendering: Efficient light transport in refractive objects. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 26(3):59–1 – 59–8, 2007.

30.10.6 Fast, Accurate, and Scalable Dynamic Height Field Rendering

Investigators: Art Tevs and Ivo Ihrke

Figure 30.50: Different height-maps rendered with Maximum Mipmap data structure [1]. All renderings are performed in real-time. Last picture shows that there are no artifacts even for narrow viewing angles.

We have developed a GPU-based, fast, and accurate dynamic height field rendering

technique that scales well to large scale height fields. Current real-time rendering algorithms for dynamic height fields employ approximate ray-height field intersection methods, whereas accurate algorithms require pre-computation in the order of seconds to minutes and are thus not suitable for dynamic height field rendering. We alleviate this problem by using *maximum mipmaps*, a hierarchical data structure supporting accurate and efficient rendering while simultaneously lowering the pre-computation costs to negligible levels. Furthermore, the technique supports view-dependent level-of-detail rendering. For the results see Figure 30.50.

References

- [1] A. Tevs, I. Ihrke, and H.-P. Seidel. Maximum mipmaps for fast, accurate, and scalable dynamic height field rendering. In *Symposium on Interactive 3D Graphics (I3D 2008)*, Redwood City, California, USA, 2008, pp. 183–190. ACM.

30.10.7 GigaVoxels: Ray-guided Streaming for Efficient Voxel Rendering

Investigator: Elmar Eisemann

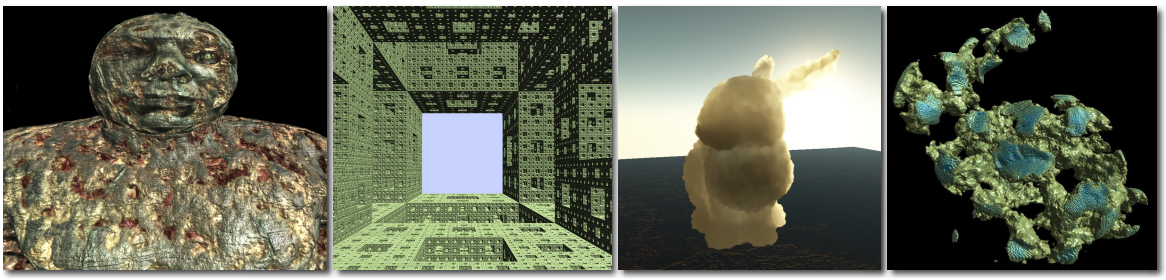


Figure 30.51: Our system [1] allows the interactive to real-time display of large volumes. This works on medical data (left) of several Gigabytes, fractals of theoretically infinite resolution (middle left), simulated data (middle right) or on-the-fly transformations (right).

We proposed a new approach to efficiently render large volumetric data sets. The system achieves interactive to real-time rendering performance for several billion voxels. Our solution is based on an adaptive data representation depending on the current view and occlusion information, coupled to an efficient ray-casting rendering algorithm. One key element of our method is to guide data production and streaming directly based on information extracted during rendering. Our work shows that volumetric models might become a valuable alternative as a rendering primitive for real-time applications. In this spirit, we allow a quality/performance trade-off and exploit temporal coherence. We also introduce a mipmapping-like process that allows for an increased display rate and better quality through high quality filtering.

References

- [1] C. Crassin, F. Neyret, S. Lefebvre, and E. Eisemann. Gigavoxels: Ray-guided streaming for efficient and detailed voxel rendering. In *ACM Symposium on Interactive 3D Graphics and Games (i3D)*, Boston, USA, 2009, pp. 15–22. ACM.

30.10.8 High-speed Marching Cubes using Histogram Pyramids

Investigator: Gernot Ziegler

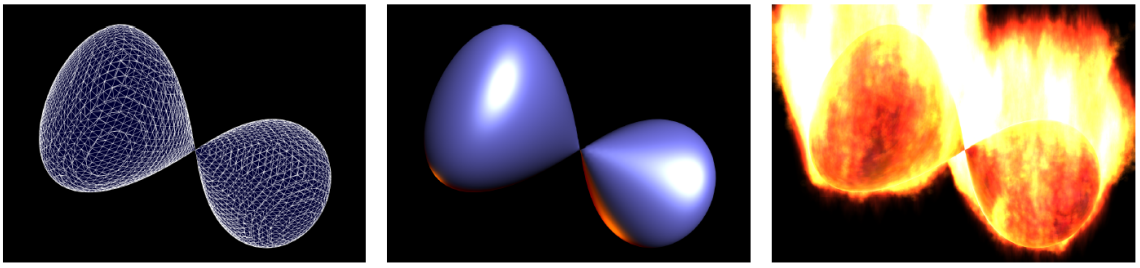


Figure 30.52: An iso-surface represented explicitly as a compact list of triangles (left) can be visualized from any viewpoint (middle) and even be directly post-processed. One example for such post-processing is the spawning of particles evenly over the surface (right). In all three images, the GPU has autonomously extracted the mesh from the scalar field, where it is kept in graphics memory.

We have developed an approach for Marching Cubes on graphics hardware for OpenGL 2.0 or comparable APIs [1]. It currently outperforms all other known GPU-based iso-surface extraction algorithms in direct rendering for sparse or large volumes, even those using the recently introduced geometry shader capabilites. To achieve this, we outfit the HistoPyramid algorithm, previously only used in GPU data compaction, with the capability for arbitrary data expansion. After reformulation of Marching Cubes as a data compaction and expansion process, the HistoPyramid algorithm becomes the core of a highly efficient and interactive Marching Cube implementation. For graphics hardware lacking geometry shaders, such as mobile GPUs, the concept of HistoPyramid data expansion is easily generalized, opening new application domains in mobile visual computing. Further, to serve recent developments, we present how the HistoPyramid can be implemented in the parallel programming language CUDA, by using a novel 1D chunk/layer construction. Some of the results can be seen on Figure 30.52.

References

- [1] C. Dyken, G. Ziegler, C. Theobalt, and H.-P. Seidel. High-speed marching cubes using histopyramids. *Computer Graphics Forum*, 27(8):2028–2039, 2008.

30.10.9 Animating Pictures of Fluid using Video Examples

Investigator: Makoto Okabe

We have developed a system that allows the user to design a continuous flow animation starting from a still fluid image [1]. The basic idea is to apply the fluid motion extracted from a video example to the target image. The system first decomposes the video example into three components, an average image, a flow field and residuals. The user then specifies equivalent information over the target image. The user manually paints the rough flow field, and the system automatically refines it using the estimated gradients of the target image. The user semi-automatically transfers the residuals onto the target image. The system then approximates the average image and synthesizes an animation on the target image by adding the transferred residuals and warping them according to the user-specified flow field. Finally, the system adjusts the appearance of the resulting animation by applying histogram matching. We designed animations of various pictures, such as rivers, waterfalls, fires, and smoke. An overview of the system is shown in Figure 30.53.

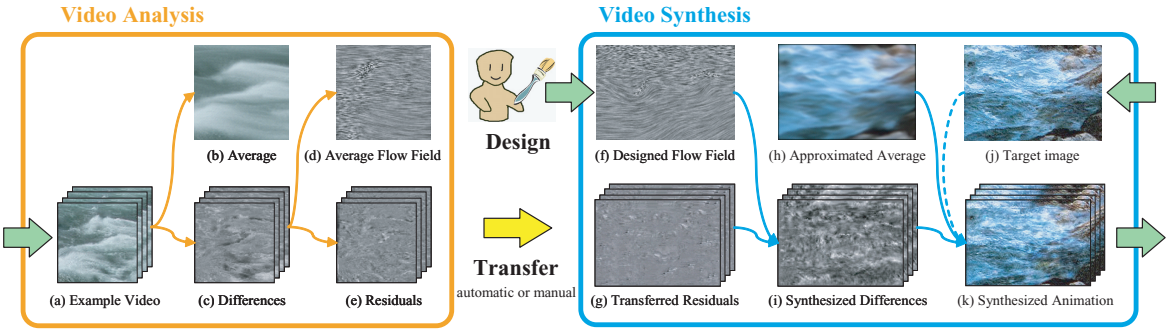


Figure 30.53: System overview. Green arrows represent input and output of our system. The video example (a) is decomposed into the average image (b) and the differences between the original frames and the average image (c). The differences are then decomposed into the flow field (d) and the residuals (e). The user semi-automatically designs the flow field (f) and transfers the residuals (g). The system then computes the approximated average image (h) by applying motion blur to the target image, an cutout from Thomas Moran's painting (j). The synthesized differences (i) are computed by combining the designed flow field and transferred residuals. Finally, to preserve the original appearance of the target image (j), the system applies histogram matching (dashed line) to compose the final flow animation (k).

References

- [1] M. Okabe, K. Anjyo, T. Igarashi, and H.-P. Seidel. Animating pictures of fluid using video examples. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28, 2009.

30.11 High Dynamic Range Imaging and Perception Issues in Graphics

Coordinator: Karol Myszkowski

The goal of this group is to develop algorithms for capturing, processing, storing and display of images and video that preserve the the appearance of original scenes. Such algorithms often take into account the performance of the human visual system to find the best compromise between fidelity and computational cost.

High dynamic range imaging (HDRI) addresses the shortcoming of the traditional imaging by capturing, processing and storing visual information with much higher precision [7]. In this group we take advantage of such high dynamic representation to develop algorithms for efficient storage [4, 5, 6] and display of high fidelity digital images and video [1, 3]. We are also interested in HDRI applications in computer graphics [2].

Real-world scenes are not only brighter and more colorful than their digital reproductions, but also contain much higher contrast, both local between neighboring objects, and global between distant objects. The eye has evolved to cope with such high contrast and its presence in a scene evokes important perceptual cues. Traditional low dynamic range (LDI) imaging, unlike HDRI, is not able to represent such high-contrast scenes. For this purpose tone mapping is employed to accommodate HDR content to LDR devices. While the overview of this important issue we have recently summarized in the book chapter [3], more recent results are presented in Section 30.11.1. Atop of tone mapping we have been interested in the problem increasing apparent (perceived) contrast on devices with limited dynamic range using the so-called Cornsweet illusion, which we discuss in Section 30.11.2. Similarly, traditional images can hardly represent common visual phenomena, such as self-luminous surfaces (sun, shining lamps) and bright specular highlights. They also do not contain enough information to reproduce visual glare (brightening of the areas surrounding shining objects). In Section 30.11.3 we discuss how to reproduce these efficiently effects on LDR displays. Finally, in Section 30.11.4 we discuss the problem of HDR image quality evaluation.

References

- [1] F. Banterle, K. Debattista, A. Artusi, S. Pattanaik, K. Myszkowski, P. Ledda, M. Bloj, and A. Chalmers. High dynamic range imaging and ldr expansion for generating hdr content. In *EUROGRAPHICS State-of-the-Art Report*, Munich, 2009, pp. 17–44. Eurographics.

- [2] M. Goesele and K. Myszkowski. Hdr applications in computer graphics. In B. Hoefflinger, ed., *High-Dynamic-Range (HDR) Vision*, Springer Series in Advanced Microelectronics (26), pp. 193–210. Springer, Heidelberg, 2007.

- [3] G. Krawczyk, K. Myszkowski, and D. Brosch. Hdr tone mapping. In B. Hoefflinger, ed., *High-Dynamic-Range (HDR) Vision*, Springer Series in Advanced Microelectronics (26), pp. 147–178. Springer, Heidelberg, 2007.

- [4] R. Mantiuk. Hdr image and video compression. In B. Hoefflinger, ed., *High-Dynamic-Range (HDR) Vision*, Springer Series in Advanced Microelectronics (26), pp. 179–192. Springer, Heidelberg, 2007.

- [5] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. High dynamic range image and video compression - fidelity matching human visual performance. In *IEEE International Conference on Image Processing (ICIP 2007)*, San Antonio, TX, USA, 2007, vol. 1, pp. 9–12. IEEE.
- [6] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Method and apparatus for encoding high dynamic range video. In *US Patent 7,483,486*. 2009.
- [7] K. Myszkowski, R. Mantiuk, and G. Krawczyk. *High Dynamic Range Video*. Synthesis Digital Library of Engineering and Computer Science. Morgan & Claypool Publishers, San Rafael, USA, 2008.

30.11.1 Tone Mapping

Investigators: Rafał Mantiuk, Grzegorz Krawczyk, Piotr Didyk, and Karol Myszkowski

Reproducing natural and artificial scenes on display devices of a limited dynamic range (contrast) is a challenging problem in photography, cinematography, printing and visualization. So far, the best results are achieved when each image is manually adjusted on the target display. This, however, is a tedious task that often requires expert skills. The question arises whether the manual adjustments can be replaced with a computational algorithm. In [3] we address this question by demonstrating that the image reproduction tasks can be formulated as an optimization problem steered by a visual metric (refer to Figure 30.54). The tone scale is strongly correlated with color appearance, which leads to disruptions in colors when tone scale is manipulated. In [4] we quantify and model the correction in color saturation that needs to be made after tone-scale manipulation. The choice of a proper exposure (brightness) for a particular image has a profound impact on an image appearance. In [2] we investigate the relation between subjectively adjusted brightness and image statistics in order to find a model for an automatic exposure adjustment. Since many tone-mapping techniques share similar concepts, in [6] we make an attempt to generalize image processing operations performed in tone mapping. The proposed generic tone mapping operator can find applications in backward-compatible HDR image and video compression, synthesis of new tone mapping operators and analysis of existing operators. Also, we are concerned with the problem of tone mapping efficiency. In [7] we propose an interactive GPU-based solver for the multi-resolution gradient domain tone mapping that we presented in [5].

The current display technologies can show much higher dynamic range (contrast), better brightness, lower black level and more saturated colors than their CRT predecessors. However, the available video content cannot take full advantage of these new capabilities. The resolution and to some extent the dynamic range of DVD movies can be improved by rescanning film negatives and color grading them for new displays. This, however, cannot restore very bright image features, such as light sources or specular highlights. We propose a semi-automatic algorithm for the restoration of bright image features in [1].

References

- [1] P. Didyk, R. Mantiuk, M. Hein, and H.-P. Seidel. Enhancement of bright video features for hdr displays. *Computer Graphics Forum*, 27(4):1265–1274, 2008.
- [2] G. Krawczyk, R. Mantiuk, D. Zdrojewska, and H.-P. Seidel. Brightness adjustment for hdr and tone mapped images. In *15th Pacific Conference on Computer Graphics and Applications*, Maui, Hawaii, 2007, pp. 373–381. IEEE.

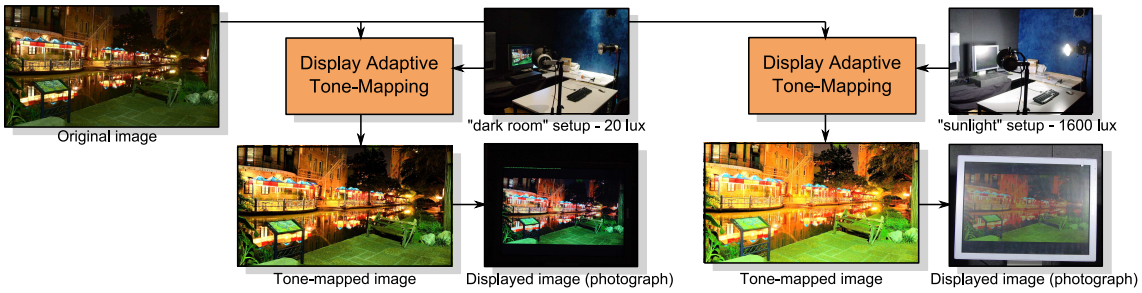


Figure 30.54: Image reproduced adaptively for low ambient light (dark room scenario – left) and high ambient light (sunlight scenario – right). The display adaptive tone mapping can account for screen reflections when generating images that optimize visible contrast.

- [3] R. Mantiuk, S. Daly, and L. Kerofsky. Display adaptive tone mapping. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 27(3):68, 2008.
- [4] R. Mantiuk, R. Mantiuk, A. Tomaszewska, and W. Heidrich. Color correction for tone mapping. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28(3), 2009.

[5] R. Mantiuk, K. Myszkowski, and H.-P. Seidel. A perceptual framework for contrast processing of high dynamic range images. *ACM Transactions on Applied Perception*, 3(3):286–308, 2006. This is a revised and extended version of the publication of the same title in the Proceedings of Second Symposium on Applied Perception in Graphics and Visualization 2005.

- [6] R. Mantiuk and H.-P. Seidel. Modeling a generic tone-mapping operator. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 27(2):699–708, 2008.
- [7] R. Mantiuk, D. Zdrojewska, A. Tomaszewska, R. Mantiuk, and K. Myszkowski. Selected problems of high dynamic range video compression and gpu-based contrast domain tone mapping. In K. Myszkowski, ed., *SCCG '08: Proceedings of the 24th Spring Conference on Computer Graphics*, Budmerice, Slovakia, 2008, pp. 11–18. ACM.

30.11.2 Contrast Enhancement using Cornsweet Illusion

Investigators: Kaleigh Smith, Grzegorz Krawczyk, Tobias Ritschel, Thorsten Grosch, and Karol Myszkowski

Contrast in photographic and computer-generated imagery communicates color and lightness differences that would be perceived when viewing the represented scene. Due to depiction constraints, the amount of displayable contrast is limited, reducing the image's ability to accurately represent the scene. Local contrast enhancement can overcome these constraints to produce improved imagery with higher information content, resulting in more efficient depictions. One such technique, called unsharp masking, influences the perception of contrast by adding high-frequency contours to an image. It is argued that local contrast enhancements made with unsharp masking can be explained by the Cornsweet illusion. This illusion creates a perceived brightening or darkening of regions that are adjacent to a specially shaped high-frequency contour, much like those that are introduced through unsharp masking. The same

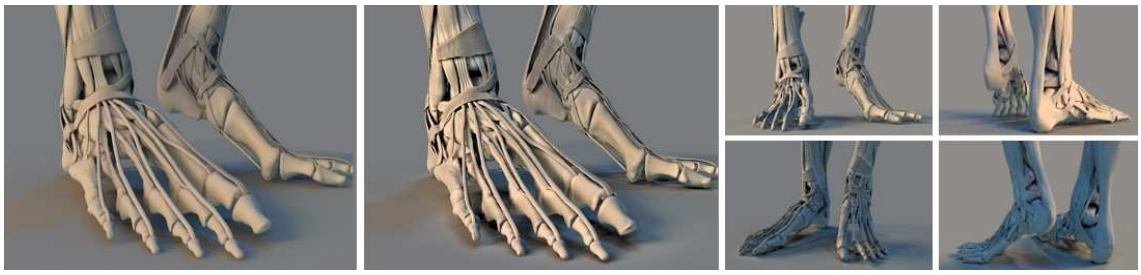


Figure 30.55: A naturally illuminated 3D scene (left) and the same scene with 3D unsharp masking enhancement (center). Our enhancement technique is coherent with the scene itself, not simply with each rendered frame, permits arbitrary lighting and is temporally coherent.

perceptual mechanism may explain why contours added to an image increase its apparent contrast.

In three novel algorithms inspired by unsharp masking, local contrast enhancement is shown to overcome a limited dynamic range [2], overcome an achromatic palette [4], and to improve the rendering of 3D shapes and scenes [3].

Image contrast reduction is common in tone mapping of high dynamic range (HDR) images (refer to Section 30.11.1). lighting. By utilizing the original HDR images as reference unsharp masking is used to restore lost contrast in tone mapped images [2]. Unsharp masking can be generalized by coupling it with a multi-resolution local contrast metric to automatically create the countershading profiles from the sub-band components which are individually adjusted to each corrected feature to best enhance contrast with respect to the reference. Additionally, a visual detection model is employed to assure that introduced contrast enhancements are not perceived as objectionable halo artifacts. The overall appearance of images remains mostly unchanged and the enhancement is achieved within the available dynamic range.

One of the most basic tools in digital image editing software is the greyscale converter, which takes a color image and produces a colorless version. The depiction challenge is to ensure that the original's chromatic contrasts are communicated even when no color is present. Apparent Greyscale [4] is a quick and simple method for converting complex images and video to perceptually accurate greyscale versions. It is a two-step approach to globally assign grey values and determine color ordering, that then locally enhances the greyscale to reproduce the original contrast. The global mapping is image independent and incorporates the Helmholtz-Kohlrausch color appearance effect for predicting differences between isoluminant colors. The multiscale local contrast enhancement reintroduces lost discontinuities only in regions that insufficiently represent original chromatic contrast. All operations are restricted so that they preserve the overall image appearance, lightness range and differences, color ordering, and spatial details, resulting in perceptually accurate achromatic reproductions of the color original.

In 3D rendering, the virtual camera settings are fixed, meaning that not all relevant scene information may be visible. The 3D Unsharp Masking approach [3] enhances local scene

contrast by unsharp masking over arbitrary surfaces under any form of illumination (refer to Figure 30.55). The adaptation of the well-known 2D unsharp masking technique to 3D interactive scenarios is designed to aid viewers in tasks like understanding complex or detailed geometric models, medical visualization and navigation in virtual environments. Its holistic approach enhances the depiction of various visual cues, including gradients from surface shading, surface reflectance, shadows, and highlights, to ease estimation of viewpoint, lighting conditions, shapes of objects and their world-space organization. The operation runs at real-time rates on a GPU and the effect is easily controlled interactively within the rendering pipeline. It is validated by psychophysical experiments [1] showing that the enhanced images are perceived as having better contrast and are preferred over unenhanced originals.

For overall contrast enhancement on modern displays that reproduce better and better black levels it becomes important to account for lower sensitivity of the human eye in dark image regions. This effect is investigated in the context of HDR displays and a correction model is proposed for standard contrast enhancement techniques [6, 5].

References

- [1] M. Ihrke, T. Ritschel, K. Smith, T. Grosch, K. Myszkowski, and H.-P. Seidel. A perceptual evaluation of 3d unsharp masking. In B. E. Rogowitz and T. N. Pappas, eds., *Human Vision and Electronic Imaging XIV, IS&T/SPIE's 21st Annual Symposium on Electronic Imaging*, San Jose, USA, 2009, *Annual Symposium on Electronic Imaging*, vol. 7240. SPIE.
- [2] G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Contrast restoration by adaptive countershading. In D. Cohen-Or and P. Slavik, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Prague, Czech Republic, 2007, vol. 26(3), pp. 581–590. Blackwell.
- [3] T. Ritschel, K. Smith, M. Ihrke, T. Grosch, K. Myszkowski, and H.-P. Seidel. 3d unsharp masking for scene coherent enhancement. In G. Turk, ed., *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, Los Angeles, USA, 2008, vol. 27(3), pp. 1–8. ACM.
- [4] K. Smith, P.-E. Landes, J. Thollot, and K. Myszkowski. Apparent greyscale: A simple and fast conversion to perceptually accurate images and video. In R. Scopigno and E. Gröller, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Crete, Greece, 2008, vol. 27(2), pp. 193–200. Blackwell.
- [5] A. Yoshida, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Perceptual uniformity of contrast scaling in complex images. In C. Wallraven and V. Sundstedt, eds., *Proceedings of ACM Symposium on Applied Perception in Graphics and Visualization*, New York, USA, 2007, pp. 137–137. ACM.
- [6] A. Yoshida, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Perception-based contrast enhancement model for complex images in high dynamic range. In B. E. Rogowitz and T. N. Pappas, eds., *Human Vision and Electronic Imaging XIII, IS&T/SPIE's 20th Annual Symposium on Electronic Imaging (2008)*, San Jose, CA, USA, 2008, vol. 6806, pp. 68060C–1–11. SPIE.

30.11.3 Brightness Enhancement Using Glare Effect

Investigators: Tobias Ritschel, Akiko Yoshida, Rafał Mantiuk, and Karol Myszkowski

The glare illusion is commonly used in CG rendering, especially in game engines, to achieve a higher brightness than that of the maximum luminance of a display. Glare is a consequence of light scattered within the human eye when looking at bright light sources. This effect can

be exploited for tone mapping since adding glare to the depiction of high-dynamic range (HDR) imagery on a low-dynamic range (LDR) medium can dramatically increase perceived contrast.

In [2] we measure the perceived luminance of the glare illusion in a psychophysical experiment. To evoke the illusion, an image is convolved with either a point spread function (PSF) of the eye or a Gaussian kernel. It has been found that 1) the Gaussian kernel evokes an illusion of the same or higher strength than that produced by the PSF while being computationally much less expensive, 2) the glare illusion can raise the perceived luminance by 20%–35%, 3) some convolution kernels can produce undesirable Mach-band effects and thereby reduce the brightness boost of the glare illusion. The reported results have practical implications for glare rendering in computer graphics.

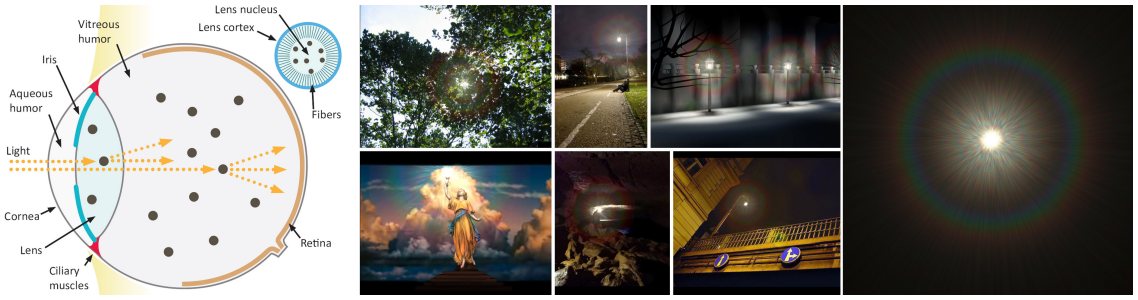


Figure 30.56: Temporal Glare [1]: Using a model of the dynamic human eye (Left), we can simulate a dynamic point spread function (Right) that can be added to images to improve perceived contrast (Middle).

Even though most, if not all, subjects report perceiving glare as a bright pattern that fluctuates in time, up to now it has only been modeled as a static phenomenon. In [1] we argued that the temporal properties of glare are a strong means to increase perceived brightness and to produce realistic and attractive renderings of bright light sources. Based on the anatomy of the human eye, we proposed a model that enables real-time simulation of dynamic glare on a GPU (refer to Figure 30.56). This allows an improved depiction of HDR images on LDR media for interactive applications like games, feature films, or even by adding movement to initially static HDR images. By conducting psychophysical studies, we validated that our method improves perceived brightness and that dynamic glare-renderings are often perceived as more attractive depending on the chosen scene.

References

- [1] T. Ritschel, M. Ihrke, J. R. Frisvad, J. Coppens, K. Myszkowski, and H.-P. Seidel. Temporal glare: Real-time dynamic simulation of the scattering in the human eye. *Computer Graphics Forum (Proc. EUROGRAPHICS 2009)*, 28(3):183–192, 2009.
- [2] A. Yoshida, M. Ihrke, R. Mantiuk, and H.-P. Seidel. Brightness of the glare illusion. In S. Creem-Regehr and K. Myszkowski, eds., *Proceedings of ACM Symposium on Applied Perception in Graphics and Visualization*, Los Angeles, CA, USA, 2008, pp. 83–90. ACM.

Figure 30.57: Objective quality evaluation of a tone mapped image (left) with respect to the HDR reference (center). The resulting distortion map (right) shows loss of visible contrast in green, and reversal of visible contrast in red. The luminance of the tone mapped image is also shown in grayscale to augment the interpretability of the map.

30.11.4 HDR Image Quality Evaluation

Investigators: Rafał Mantiuk, Tunç O. Aydın, and Karol Myszkowski

Is it possible to *compute* the rate of quality of a distorted test image with respect to a reference image without human involvement? Image quality assessment metrics are commonly used to predict the outcome of costly subjective image quality studies. We addressed three shortcomings of current metrics: the lack of HDR capability, the inability to work with a test–reference image pair with different dynamic ranges, and the missing components to make predictions under temporally lighting conditions.

Widely used image quality metrics such as MSE, PSNR and SSIM take as input gamma corrected images. In their computation, it is implicitly assumed that these pixel values are perceptually uniform. While this is approximately the case for typical CRT displays (from 0.1 to 80 cd/m^2), it is no longer true for much brighter LCD displays (up to 500 cd/m^2), plasma displays (small regions up to 1000 cd/m^2) and experimental HDR displays (up to 3000 cd/m^2). To predict the quality of images shown on bright displays, we developed a straightforward extension to PSNR and SSIM, that makes them capable of handling all luminance levels visible to the human eye [2]. The extension consist of a "perceptually uniform" encoding whose response is optimized to fit sRGB response within the dynamic range of a CRT display.

A more challenging problem is the objective prediction of image quality, where the reference and test images have different dynamic ranges. We developed a novel image quality metric that detects "structural changes" in the test image utilizing a HDR capable human visual system model [1]. Structural changes are detected through three distortion measures: loss of visible contrast, amplification of invisible contrast, and reversal of visible contrast. The outcome is visualized in form of an in-context distortion map (Figure 30.57). The metric is also validated through a subjective study involving 14 subjects, where we show the subjective responses highly correlate with the computed quality predictions. Applications of the metric include objective evaluation of tone mapping operators and quality evaluation on displays with different capabilities.

Image quality studies often assume a laboratory setting with highly controlled lighting, and ignore the real world conditions such as dynamically changing illumination and resulting temporally changing adaptation states of the visual system. We extend [1] by removing the assumption that the visual system is perfectly adapted to the image that has been evaluated, with a temporal adaptation model that predicts the degree of *maladaptation* at each instance [3]. We embed our model to a lighting simulator that computes the illumination inside a car. The simulator also includes a precisely modeled interior display utilizing actual display BRDF data. The resulting system is used to predict the near- and supra-threshold visibility of the display content under various lighting conditions and in the presence of reflections.

Finally, objective image quality assessment methods are used in medical imaging context to predict the impairment caused by JPEG 2000 compression artifacts on chest and abdominal CT scans [5, 4, 6].

References

- [1] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Dynamic range independent image quality assessment. In G. Turk, ed., *Proceedings of ACM SIGGRAPH 2008*, Los Angeles, USA, 2008, *ACM Transactions on Graphics*, vol. 27(3), pp. 1–10. ACM.

- [2] T. O. Aydin, R. Mantiuk, and H.-P. Seidel. Extending quality metrics to full luminance range images. In A. Heimer, ed., *Human Vision and Electronic Imaging XIII*, San Jose, USA, 2008. SPIE.

- [3] T. O. Aydin, K. Myszkowski, and H.-P. Seidel. Predicting display visibility under dynamically changing lighting conditions. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28(2):173–182, 2009.

- [4] B. Kim, K. H. Lee, K. J. Kim, R. Mantiuk, S. Hahn, T. J. Kim, and Y. H. Kim. Prediction of perceptible artifacts in jpeg 2000-compressed chest ct images using mathematical and perceptual quality metrics. *American Journal of Roentgenology*, 190:328–334, 2008.

- [5] B. Kim, K. H. Lee, K. J. Kim, R. Mantiuk, H.-r. Kim, and Y. H. Kim. Artifacts in slab average-intensity-projection images reformatted from jpeg 2000 compressed thin-section abdominal ct data sets. *American Journal of Roentgenology*, 190:W342–W350, 2008.

- [6] K. J. Kim, B. Kim, K. H. Lee, T. J. Kim, R. Mantiuk, H.-S. Kang, and Y. H. Kim. Regional difference in compression artifacts in low-dose chest ct images: Effects of mathematical and perceptual factors. *American Journal of Roentgenology*, 191:W30–W37, 2008.

30.12 Software

As part of the research process, several libraries, development tools, and application frameworks have been developed by members of the group. In this section we describe some of them that evolved to a level where it was appropriate to either distribute them as *open source* projects or let members of other research institutions benefit from software that had been developed in our group.

30.12.1 PFSTOOLS for Processing High Dynamic Range Images and Video

Investigators: Rafał Mantiuk, Grzegorz Krawczyk, and Tunç O. Aydın

The `pfstools` package is a set of command line programs for reading, writing, manipulating and viewing high-dynamic range (HDR) images and video frames. All programs in the package exchange data using a simple generic high dynamic range image format, `pfs`, and they use unix pipes to pass data between programs and to construct complex image processing operations.

`pfstools` come with a library for reading and writing pfs files. The library can be used for writing custom applications that can integrate with the existing `pfstools` programs.

`pfstools` offer also a good integration with a high-level mathematical programming languages, such as MATLAB or GNU Octave. `pfstools` can be used as the extension of MATLAB or Octave for reading and writing HDR images or simply to store effectively large matrices.

The `pfstools` package is an attempt to integrate the existing high dynamic range image formats by providing a simple data format that can be used to exchange data between applications.

The `pfstools` package is accompanied by `pfscalibration` and `pfstmo` packages. The `pfscalibration` package provides an algorithm for the photometric calibration of cameras and for the recovery of high dynamic range (HDR) images from the set of low dynamic range (LDR) exposures. The `pfstmo` package contains the implementation of seven state-of-the-art tone mapping operators suitable for convenient processing of both static images and animations.

The `pfstools`, `pfscalibration` and `pfstmo` packages are licensed as an Open Source project under a General Public License (GPL). The project web pages can be found at:

`http://www.mpi-inf.mpg.de/resources/pfstools/`

`http://www.mpi-inf.mpg.de/resources/hdr/calibration/pfs.html`

`http://www.mpi-inf.mpg.de/resources/tmo/`

The software was extensively used and tested within the scope of all projects described in Section 30.11. The software received wider interest of Open Source community and third party contributors prepared installation packages which are now included in several Linux distributions including Debian, Fedora and Suse. The software was presented on the *Electronic Imaging Conference 2007* and a general introduction to the package was published in the proceedings [1].

References

- [1] R. Mantiuk, G. Krawczyk, R. Mantiuk, and H.-P. Seidel. High dynamic range imaging pipeline: Perception-motivated representation of visual content. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., *Human Vision and Electronic Imaging XII*, San Jose, CA, USA, 2007, *SPIE*, vol. 6492, p. 649212. SPIE.

30.12.2 Image Quality Assessment Online

Investigator: Tunç O. Aydın

Modern image quality assessment methods involving models of early human vision are often cumbersome to implement and use for computer graphics and computer vision researchers who are not specialized in visual perception. The steep learning curve of these tools is a significant obstacle for researchers who are only interested to evaluate the results of their image processing methods. In order to make state of the art image quality assessment methods accessible to a wide audience of researchers, we implemented a web service where one can assess the quality of a *distorted* image with respect to a *reference* image. The service is located at:

`http://drim.mpi-inf.mpg.de`

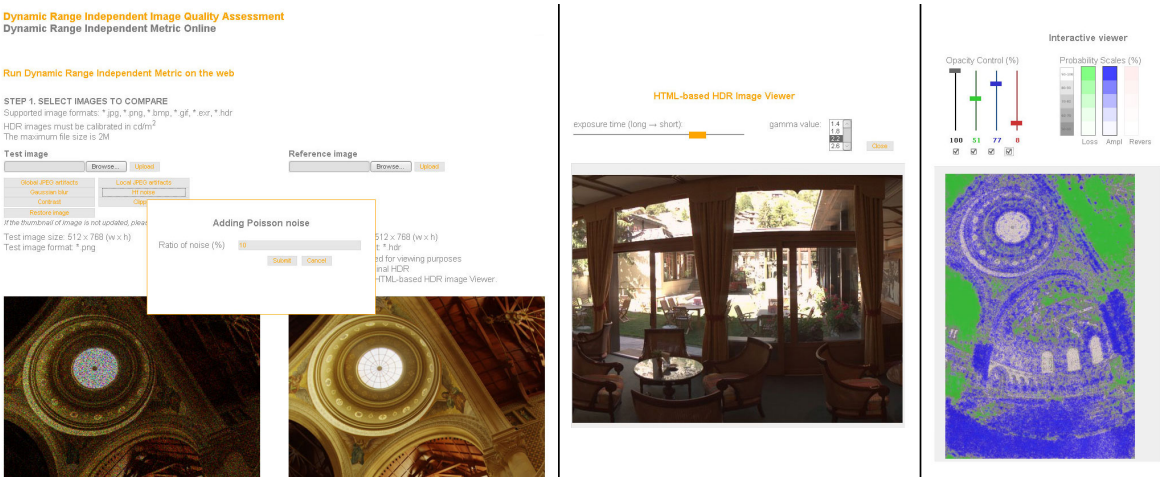


Figure 30.58: Left: The user uploads the distorted and reference images. The distorted image can also be generated by manipulating the reference using the built-in distortion generator. Center: The HTML based viewer enables viewing HDR images at various exposures and gamma values. Right: Visualization of the distortion map generated by the dynamic range independent metric. Contrast loss and amplification are shown in green and blue, respectively. The interface allows setting the opacity for each distortion type. In this example, contrast reversal is set to a very low value to prevent the occlusion of other distortions.

In the first step, the user uploads the distorted and reference images (Figure 30.58 left). Currently, our service supports major 8-bit (JPEG, PNG, BMP, GIF) and HDR (Radiance HDR, EXR) image formats. HDR images are tone mapped to provide a preview. A better visualization is provided by our HTML based HDR image viewer tool, where one can dynamically adjust the exposure (Figure 30.58 center). The uploaded image pair is then processed by one or more of the three methods: PSNR can be used for a simple and fast comparison, HDR-VDP models early vision pipeline and supports HDR images in addition to LDR images [2], and finally the dynamic range independent metric enables quality assessment

of LDR images with respect to HDR images, and vice versa [1]. The user can set viewing distance, pixels per visual degree, and peak contrast values for the latter two methods. Similarly for LDR images, the display's gamma correction and maximum luminance can be set through the online form. For convenience, we provide default values and presets of all parameters. After the processing is complete, the resulting distortion maps are visualized using the interactive viewer (Figure 30.58 right). If the user chooses multiple quality metrics, results for all metrics are evaluated and shown side-by-side for comparison.

References

- [1] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Dynamic range independent image quality assessment. In G. Turk, ed., *Proceedings of ACM SIGGRAPH 2008*, Los Angeles, USA, 2008, *ACM Transactions on Graphics*, vol. 27(3), pp. 1–10. ACM.

[2] R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Visible difference predictor for high dynamic range images. In W. Thissen, P. Wieringa, M. Pantic, and M. Ludema, eds., *2004 IEEE International Conference on Systems, Man & Cybernetics*, Hague, The Netherlands, 2004, vol. 3, pp. 2763–2769. IEEE.

30.12.3 Plexus

Investigator: Tobias Ritschel

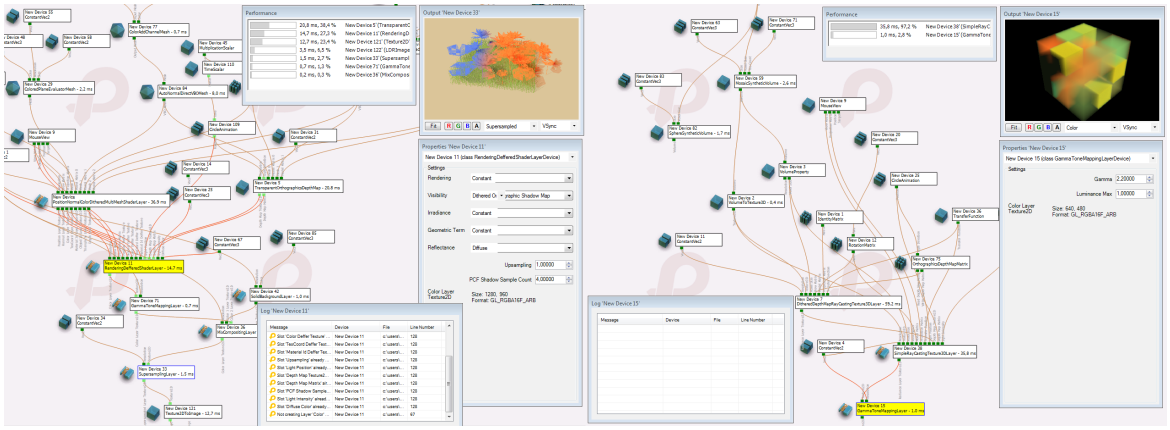


Figure 30.59: Two screen-shots form ,,Plexus", showing a typical use-case where an experimental shadow algorithm is tested on both surface- (left) and volume (right) rendering.

Plexus is a software system provided by Tobias Ritschel to design, implement and test algorithms for visual computing based on a visual dataflow paradigm. Dataflow occurs in graphs containing connected devices, which are manipulated using a visual editor at runtime and serialized using XML. Many such devices are implemented using recent graphics hardware, but the system is not limited to that purpose. While many concurrent systems in visual

computing are specialized on dataflow using one kind of data (matrices, images, meshes, volumes, hardware shading fragments, etc.), Plexus is more general and handles all of them. One key objective is to make instrumentation of visual computing easier: all devices can be inspected (e.g. their time and memory consumption) and manipulated visually with a few clicks, easing discussion, understanding and tuning of results. It implemented in (mostly platform-independent) C++, CUDA and GLSL (112 kLOC), with a user interface written in managed C++ for Microsoft Windows. The system so far has only been made available to selected research groups, but will become open source at a later point in time.

30.12.4 XGRT – Extensible Graphics Toolkit

Investigator: Michael Wand

The Extensible Graphics Toolkit (XGRT) is an object oriented platform for research in rendering and geometry processing. It consists of four main layers:

- **The foundation library.** This module adds object oriented techniques such as meta classes, structural reflection, and object serialization to C++. With this framework, it is easy to work with objects in an interactive environment. In particular, the framework provides automatic creation of user interfaces from object properties, storage of objects in files and transmission across networks as well as an object oriented scripting interface.

- **Abstract geometry containers.** A second component is a set of classes for encapsulating geometric primitives such as volume, point cloud or triangle mesh data. A particular focus is on abstract data handling – the application is unaware of the actual data layout, which might be stored out-of-core or provide additional information such as multi-resolution representations or auxiliary data structures such as spatial hierarchies for fast range queries. The data is accessed via abstract iterator objects that can be composed in a functional programming style. Employing a block-wise processing paradigm, we still obtain high processing performance, despite the high-level interface. Similarly, the data layout (for example vertex attributes such as positions, normals, colors) is kept flexible and determined at runtime, using a programmable shader interface for runtime mapping for rendering.

- **Large data sets.** In particular, we have implemented the fully dynamic hierarchical multi-resolution data structures described in section 30.4.10 as one geometry container. This data container provides multi-resolution access to very large 3D scanner data sets, with size only limited by hard drive space. It is used as basis for real-time rendering, editing and externally efficient geometry processing algorithms.

- **Interactive Editor.** At the highest logical level, the actual application is running. The system currently provides an editor for 3D point clouds with interactive tools for data touch up, processing and reconstruction. Thanks to the dynamic multi-resolution data structures, the editor is able to interactively edit extremely large data sets (we have tested it with data set sizes of up to 64 GB, which was the largest data set available to us so far). Currently, this is the only software system we are aware of that provides such interactive capabilities on out-of-core data.

Figure 30.60: The extensible graphics toolkit. From left to right: the application logo, interactive shader design for curvature analysis, procedural shading, real-time editing of large data sets, up to several GB in size.

The software is available as open source at `http://www.mpi-inf.mpg.de/~mwand/XGRT`. Figure 30.60 shows some screenshots of the toolkit in different application scenarios.

30.12.5 osgPPU - OpenSceneGraph NodeKit

Investigator: Art Tevs

Figure 30.61: Screenshots from the video created for [1]. The use of various post processing effects achieved with our NodeKit increase the visual appealing of the shown results dramatically.

osgPPU is a NodeKit to use with OpenSceneGraph a open source scene graph library `http://www.openscenegraph.org`. It provides a graph based specification of a computation pipeline which is based on so called PostProcessingUnits (PPUs). Each ppu does render a screen aligned quad in a frame buffer object. During the rendering a shader can be applied. The results (there could be many per one pass) are passed to the next ppu in the graph. The outcoming result of the pipeline can either be shown on the screen or used as a texture for other further processing.

The library was first used during the development of [1] and [2], where it has been used to create various post processing effects, like HDR-Rendering, Depth Of Field, Motion Blur, Tone Mapping etc. (see Figure 30.61). Furthermore the library can also be used for some off-line computation tasks, which has to be performed on the GPU. In [3] we have used the library to compute internal data structures of the algorithm, which then ran on the GPU to compute the results. osgPPU is implemented completely platform independent. It is programmed in C++ and was successfully tested in Linux and Windows environments. The osgPPU NodeKit is an open source library which is available to the public audience at `http://projects.tevs.eu/osgppu`. It has been successfully included into various applications which are based on OpenSceneGraph and the number of users using the nodekit is increasing continuously.

References

- [1] I. Ihrke, G. Ziegler, A. Tevs, C. Theobalt, M. Magnor, and H.-P. Seidel. Eikonal rendering: Efficient light transport in refractive objects. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 26(3):59–1 – 59–8, 2007.
- [2] A. Tevs. Realistic real-time rendering of refractive objects. Masters thesis, Universität des Saarlandes, Saarbruecken, Germany, 2007.
- [3] A. Tevs, I. Ihrke, and H.-P. Seidel. Maximum mipmaps for fast, accurate, and scalable dynamic height field rendering. In *Symposium on Interactive 3D Graphics (I3D 2008)*, Redwood City, California, USA, 2008, pp. 183–190. ACM.

30.13 Academic Activities

30.13.1 Journal Positions

Karol Myszkowski is on the editorial board of

- *Journal of Virtual Reality and Broadcasting* (since 2004),
- *ACM Transactions on Applied Perception* (since 2002),
- *Machine Graphics & Vision* (since 1998).

Hans-Peter Seidel is on the editorial board of

- *IEEE Transactions on Visualization and Computer Graphics (IEEE TVCG)* (since 2004),
- *International Journal of Shape Modeling (IJSM)* (since 2001),
- *The Visual Computer (TVC)* (since 1999),
- *Computer Aided Geometric Design (CAGD)* (since 1999),
- *Graphical Models (GMOD)* (since 1995),
- *Computer Graphics Forum (CGF)* (since 1993).

30.13.2 Conference and Workshop Positions

Elmar Eisemann:

- *Eurographics Symposium on Rendering*, Girona, June 2009.

Hendrik Lensch:

- *Eurographics Symposium on Rendering*, Sarajevo, June 2008.
- *5th ACM/IEEE International Workshop on Projector-Camera Systems*, Los Angeles, August 2008.
- *ACM SIGGRAPH*, Los Angeles, August 2008.
- *Eurographics*, Crete, April 2008.
- *EG - International Conference on Computer Graphics Theory and Application (TPCG)*, Manchester, June 2008.
- *International Conference on Visualization, Imaging and Image Processing (VIIP)*, Palma de Mallorca, August 2007.
- *Vision, Modeling, and Visualization (VMV)*, Saarbrücken, November 2007 (Program Chair).
- *ACM SIGGRAPH*, San Diego, August 2007.
- *Eurographics Symposium on Rendering*, Grenoble, June 2007.
- *Mirage*, Rocequencourt, March 2007.

Meinard Müller:

- *2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation*, In conjunction with ICCV 2007, Rio de Janeiro, October 2007.
- *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, October 2007.
- *MCM International Conference Mathematics and Computation*, New Haven, June 2009.

Karol Myszkowski:

- *Eurographics Symposium on Rendering*, Girona, June 2009.
- *International Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging (CAe'09)*, Victoria, British Columbia, May 2009.
- *Winter School on Computer Graphics (WSCG'09)*, Plzen, February 2009.
- *Sixth International Conference on Virtual Reality, Computer Graphics, Visualisation and Interaction in Africa (Afrigraph'09)*, Pretoria, February 2009.
- *Human Vision and Electronic Imaging XIV (HVEI'09)*, San Jose, January 2009.
- *International Conference on Computer Vision and Graphics (ICCVG 2008)*, Warsaw, November 2008.
- *ACM Siggraph*, Los Angeles, August 2008.

- *ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization (APGV'08)*, Los Angeles, August 2008 (Program Co-chair).
- *IEEE Symposium on Interactive Ray Tracing*, Los Angeles, August 2008.
- *ACM Siggraph Symposium on Non-photorealistic Animation and Rendering (NPAR'08)*, Annency, June 2008.
- *Graphicon*, Moscow, June 2008.
- *Eurographics Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging (CAe'08)*, Lisbon, June 2008.
- *Spring Conference on Computer Graphics (SCCG'08)*, Budmerice Castle, April 2008 (Conference & Program Chair).
- *Winter School on Computer Graphics (WSCG'08)*, Plzen, February 2008.
- *Human Vision and Electronic Imaging XIII (HVEI'08)*, San Jose, January 2008.
- *Fifth International Conference on Computer Graphics and Interactive Techniques in Australasia and South-East Asia (GRAPHITE'07)* Kuala Lumpur, November 2007.
- *International Conference on Cyberworlds 2007* Hannover, October 2007.
- *Fourth International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa (Afrigraph'07)*, Grahamstown, October 2007.
- *Eurographics 2007*, Prague, September 2007.
- *Eurographics 2007*, Prague, September 2007 (Tutorial Co-chair)
- *ACM Siggraph Symposium on Non-photorealistic Animation and Rendering (NPAR'07)*, San Diego, August 2007.
- *ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization 2007*, Tuebingen, July 2007.
- *Eurographics Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging (CAe'07)*, Banff, June 2007.
- *Graphicon*, Moscow, June 2007.

Bodo Rosenhahn:

- *12th International Workshop on Combinatorial Image Analysis (IWCIA)*, Mexico, November 2009.
- *Vision Modeling and Visualization (VMV)*, Braunschweig, November 2009.
- *Computer Analysis of Images and Patterns (CAIP)*, Münster, September 2009.
- *Annual Symposium German Association for Pattern Recognition (DAGM)*, Jena, September 2009.
- *Australasian Joint Conference on Artificial Intelligence (AI)*, Auckland, December 2008.
- *Image and Vision Computing New Zealand (IVCNZ)*, Christchurch, November 2008.
- *European Conference on Computer Vision 2008 (ECCV)*, Marseille, October 2008.
- *Statistical and Geometrical Approaches to Visual Motion Analysis*, Castle Dagstuhl, July 2008.

- *Technical Program Vice Chair: IEEE International Conference on Multimedia and Expo (ICME)*, Hannover, June 2008.
- *12th International Workshop on Combinatorial Image Analysis (IWCIA)*, Buffalo, April 2008.
- *Computer Graphics, Visualization and Computer Vision (WSCG)*, Plzen, February 2008.
- *Robot Vision*, Auckland, February 2008.
- *ICCV Workshop Human Motion – Understanding, Modeling, Capture and Animation II*, Organizer, October 2007.

Hans-Peter Seidel:

- *ACM SIGGRAPH Asia 2009*, Papers Advisory Committee, Yokohama, December 2009.
- *SIAM/ACM Joint Conference on Geometric and Physical Modeling 2009*, San Francisco, October 2009.
- *Pacific Graphics 2009 (PG'09)*, Honorary Conference Co-Chair, Jeju, October 2009.
- *Pacific Graphics 2009 (PG'09)*, Jeju, October 2009.
- *International Conference on Computer Vision 2009 (ICCV'09)*, Kyoto, September 2009.
- *International Workshop on 3D Imaging and Modeling 2009 (3DIM'09)*, Kyoto, September 2009.
- *Shape Modeling International 2009 (SMI'09)*, Beijing, June 2009.
- *IEEE Conference on Computer Vision and Pattern Recognition 2009 (CVPR'09)*, Miami Beach, June 2009.
- *Eurographics 2009*, Munich, March 2009.
- *ACM SIGGRAPH Asia 2008*, Papers Advisory Committee, Singapore, December 2008.
- *Kickoff Workshop, Cluster of Excellence on Multimodal Computing and Interaction*, Scientific Coordinator, Saarbrücken, November 2008.
- *European Conference on Computer Vision (ECCV'08)*, Marseille, October 2008.
- *Pacific Graphics 2008 (PG'08)*, Tokyo, October 2008.
- *Workshop on 3D Face Processing 2008 (3DFP'08)*, Anchorage, June 2008.
- *International Symposium on 3D Data Processing*, Visualization and Transmission 2008 (3DPVT'08), Atlanta, June 2008.
- *Eurographics/ ACM SIGGRAPH Symposium on Geometry Processing 2008 (SGP'08)*, Copenhagen, June 2008.
- *Pacific Graphics 2007 (PG'07)*, Maui, Hawaii, November 2007.
- *Eurographics 2007*, Prague, September 2007.
- *ACM SIGGRAPH/Eurographics Symposium on Computer Animation 2007 (SCA'07)*, San Diego, August 2007.
- *Eurographics/ACM SIGGRAPH Symposium on Geometry Processing 2007 (SGP'07)*, Barcelona, July 2007.

- *Shape Modeling International 2007 (SMI'07)*, Lyon, France, June 2007.
- *ACM Symposium on Solid and Physical Modeling 2007 (SPM'07)*, Beijing, June 2007.
- *Dagstuhl Seminar on Visual Computing - Convergence of Computer Graphics and Computer Vision*, Organizer, April 2007.

Mike Sips:

- *Vision, Modeling, and Visualization Workshop*, Braunschweig, November 2009.
- *5th International Symposium on Visual Computing*, Las Vegas, November 2009.
- *IEEE VisWeek*, Organizing Committee, Atlantic City, October 2009.
- *Symposium of Information Visualization in Biomedical Informatics*, Barcelona, July 2009.
- *ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery (VAKD '09)*, Organizing Committee, Paris, June 2009.
- *Eurographics/IEEE Symposium on Visualization*, Berlin, June 2009.
- *IEEE International Conference on Data Mining*, Pisa, December 2008.
- *4th International Symposium on Visual Computing*, Las Vegas, November 2008.
- *IEEE Information Visualization Conference*, Columbus, October 2008.
- *IEEE Symposium on Visual Analytics Science and Technology*, Columbus, October 2008.
- *IEEE VisWeek*, Organizing Committee, Columbus, October 2008.
- *Vision, Modeling, and Visualization Workshop*, Saarbrücken, November 2007.
- *IEEE Symposium on Visual Analytics Science and Technology*, Sacramento, October 2007.
- *IEEE Information Visualization Conference*, Sacramento, October 2007.
- *Symposium of Information Visualization in Biomedical Informatics*, Zurich, July 2007.

Robert Strzodka:

- *IEEE 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Potsdam, May 2009.
- *Eurographics*, Munich, March 2009.
- *IEEE/ACM Workshop New Frontiers in High-performance and Hardware-aware Computing*, in conjunction with Micro 2008, Lake Como, November 2008.
- *ACM Siggraph/Eurographics Symposium on Graphics Hardware*, Sarajevo, June, 2008.
- *ACM Siggraph/Eurographics Symposium on Graphics Hardware*, San Diego, August 2007.

Thorsten Thormählen:

- *Vision, Modeling, and Visualization (VMV) 2009*, Braunschweig, November 2009.
- *Image and Vision Computing New Zealand*, Palmerston North, New Zealand, November 2009

- *European Conference on Visual Media Production*, London, November 2008
- *Image and Vision Computing New Zealand*, Christchurch, November 2008
- *European Conference on Visual Media Production*, London, November 2007

Michael Wand:

- *Vision, Modeling, and Visualization (VMV) 2009*, Braunschweig, November 2009.
- *Afrigraph 2009*, Pretoria, February 2009.
- *Symposium on Point-Based Graphics (PBG) 2008*, Los Angeles, CA, August 2008.
- *Symposium on Geometry Processing (SGP) 2008*, Copenhagen, July 2008.
- *Vision, Modeling, and Visualization (VMV) 2007*, Saarbrücken, November 2007.
- *Afrigraph 2007*, Grahamstown, October 2007.
- *Symposium on Point-Based Graphics (PBG) 2007*, Prague, September 2007.

30.13.3 Invited Talks and Tutorials

Martin Fuchs:

- *Selected Topics on Relighting in Image Space*, Talk at Princeton University, August 2008.
- *Selected Topics on Relighting in Image Space*, Talk at Siegen University, July 2008.

Elmar Eisemann:

- *Optimized Representations for the Acceleration of Display- and Collision Queries*, Invited Talk, Technische Universität Wien, March 2009.
- *Accelerated Rendering using Adapted GPU Structures*, Invited Talk, Massachusetts Institute of Technology, August 2008.
- *Fast Visibility Sampling*, Invited Talk, Massachusetts Institute of Technology, August 2007.

Zhao Dong:

- *Convolution Soft Shadow Maps and its applications*, Talk at Computer Graphics group of Microsoft Research Redmond, October 2008.

Thorsten Grosch:

- *Interactive Global Illumination on the GPU*, Talk at University of Koblenz-Landau, Campus Koblenz, July 2008.

Hendrik Lensch:

- *Principles of Appearance Acquisition and Representation*, Tutorial, ACM SIGGRAPH, August 2008.
- *Projectors for Graphics*, Tutorial, ACM SIGGRAPH, August 2008.

- *Principles of Appearance Acquisition and Representation*, Tutorial, ICCV, October 2007.
- *Capturing Reflectance – From Theory to Practice*, Tutorial, Eurographics, September 2007.
- *Acquisition and Modeling of Global Light Transport using Reflectance Fields*, Keynote Presentation, Theory and Practice of Computer Graphics, Bangor, June 2007.

Meinard Müller:

- *Analysis and Retrieval Techniques for Music and Motion Data*, Tutorial, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, April 2009.
- *Analysis and Retrieval Techniques for Motion and Music Data*, Tutorial, Eurographics (EG), Munich, March 2009.
- *Variationen über Happy Birthday – 60 Aspekte zur automatisierten Musikdatenerschließung*, Invited Talk, Festkolloquium anlässlich des 60. Geburtstags von Prof. Dr. Michael Clausen, Bonn, November 2008.
- *Multimedia Information Retrieval*, Tutorial Herbstschule Information Retrieval, Gesellschaft für Informatik, Dagstuhl, September 2008.
- *Analysis and Retrieval Techniques for Motion and Music Data*, Tutorial IEEE International Conference on Multimedia & Expo (ICME), Hannover, June 2008.
- *Content-based Multimedia Retrieval—Music and Motion*, Keynote Talk, Conference on Visual Media Production (CVMP), London, November 2007.
- *Alignment Techniques and its Applications to Information Retrieval*, Tutorial, DAAD Summer School 2007, Methods from Mathematics and Computer Science for Pattern Recognition in Biology, Shanghai, October 2007.
- *Synchronization and Matching Techniques for Music Data*, Tutorial, International Symposium on Music Information Retrieval (ISMIR), Vienna, September 2007.

Karol Myszkowski:

- *When are HDR Images better than Conventional Images?*, panel discussion member, Human Vision and Electronic Imaging XIV, San Jose, January 2009.
- *Tone Mapping: Contrast Distortion Metrics and Restoration Techniques*, Czech Technical University in Prague, Czech Republic, June 2008.
- *Tone Mapping: Contrast Distortion Metrics and Restoration Techniques*, Keynote talk, Spring Conference on Computer Graphics (SCCG'08), Budmerice, April 2008.
- *High Dynamic Range Imaging: Tone Mapping, Distortion Metrics and Restoration Techniques*, Sharp Laboratories of America, Portland, January 2008.
- *Contrast Restoration by Adaptive Countershading*, Dagstuhl Seminar on Visual Computing: Convergence of Computer Graphics and Computer Vision, April 2007.

Thomas Schultz:

- *Higher-Order Tensors in Diffusion Imaging and Image Processing*, University of Kaiserslautern, January 2009.
- *Crease Surfaces: From Theory to Extraction and Application to DT-MRI*, VRVis, Vienna, January 2009.

Hans-Peter Seidel:

- International Conference on Curves and Surfaces, Avignon, June 2009.
- Multimodal Computing and Interaction, GI-Jahrestagung, Lübeck, September 2009.
- Multimodal Computing and Interaction, Univ. Tübingen, July 2009.
- Multimodal Computing and Interaction, Univ. Koblenz, January 2009.
- Statistical and Geometrical Approaches to Visual Motion, Dagstuhl, July 2008.
- European Conference on Visual Media Production, London, November 2007.
- 12th IMA Conference on Mathematics of Surfaces, Sheffield, September 2007.
- 6th International Conference on 3D Digital Imaging and Modeling 2007, (3DIM'07), Montreal, August 2007.
- Theory and Practice of Computer Graphics, 25th Eurographics UK Conference, Bangor, June 2007.

Robert Strzodka:

- *Scientific Computing on Multi-GPU Systems*, Invited Talk, Conference on Applications of Graphics Processors in High Performance Computing, Warsaw, March 2009.
- *Mixed Precision Methods*, Invited Talk, NVISION Conference, San Jose, August 2008.
- *GPGPU and CUDA Tutorials*, Tutorial, International Conference on Architecture of Computing Systems, Dresden, February 2008.
- *Parallel Computing with Graphics Processors*, Tutorial, simula research laboratory, Oslo, May 2007.
- *Parallel Particle Level Set Method on the GPU*, Invited Talk, Simon Frasier University, Vancouver, May 2007.
- *Hybrid Computing on GPU Clusters*, Invited Talk, Los Alamos National Lab, Los Alamos, April 2007.

Michael Wand:

- *Geometric Correspondence Problems*, AI meets Graphics Seminar Series, Stanford University, February 2009.
- *Statistical Geometry Processing*, Oberseminar Theoretische Informatik, Paderborn University, November 2008.
- *Animation Reconstruction*, University of Tübingen, WSI/GRIS Colloquium, February 2008.

- *Animation Reconstruction*, Technical University Dresden, Computer Graphics Group, December 2007.
- *What do you need to scan your city?*, Invited Talks at HP Research Palo Alto, Adobe Research San Jose, and Bosch Research Palo Alto, April 2007.

30.13.4 Other Academic Activities

Karol Myszkowski:

- Steering Committee for Eurographics Symposium on Rendering (since 2008),
- Review Panel for the EU-sponsored CROSSMOD project IST-014891-2 (2006-2008).

Hans-Peter Seidel:

- Scientific Coordinator, Cluster of Excellence on Multimodal Computing and Interaction, Saarland University (since 2007),
- Governance Board, Intel Visual Computing Institute (since 2009),
- Chair, Scientific Advisory Board, Minerva Leibniz Center, Hebrew University (since 2005),
- Founding Chair, Eurographics Awards Programme (since 2004),
- Elected Member DFG Fachkollegium Informatik (German Research Council) (since 2004),
- Co-Director, Max Planck Center for Visual Computing and Communication Stanford / Saarbrücken (since 2003),
- Executive Committee, Solid Modeling Association (Steering Committee ACM Solid Modeling Symposium) (since 2003),
- Executive Committee, Eurographics Association (since 1992).

30.14 Teaching Activities

Summer Semester 2007

Courses:

Realistic Image Synthesis (P. Slusallek, K. Myszkowski)

Computational Photography (H. Lensch)

Winter Semester 2007/2008

Courses:

Computer Graphics (H. Lensch)

Summer Semester 2008

Courses:

Realistic Image Synthesis (P. Slusallek, K. Myszkowski)
Geometric Modeling (H.-P. Seidel, M. Wand)
Information Retrieval for Music and Motion (M. Müller)

Seminars:

Image-based 3D Analysis (H. Lensch, T. Thormählen)

Winter Semester 2008/2009

Courses:

Massively Parallel Computing with CUDA (H. Lensch, R. Strzodka)

Seminars:

Music Information Retrieval (M. Müller)
Information Visualization (M. Sips)

Summer Semester 2009

Courses:

Realistic Image Synthesis (P. Slusallek, K. Myszkowski)
Music Processing (M. Müller)

Seminars:

Computer Vision and Visual Special Effects (T. Thormählen, E. Eisemann)
Geometric Correspondence Problems (M. Wand)

Diploma Theses

Wenxiang Ying: Edge Based HDR Compression, 2007.
Christian Weber: Lucky Exposure Imaging Applied to Small Aperture Telescopes, 2007.
Lukas Heidenreich: Real-time Hierarchical Stereo Matching on Graphics Hardware, 2007.
Danyi Wang: 3D Shape Complexity Using View Similaritytitel, 2008.

Master Theses

Art Tevs: Realistic Real-Time Rendering of Refractive Objects, 2007.
Sascha El-Abed: Hole Filling in Images and in Video Sequence, 2007.
Kristina Scherbaum: Face Recognition and Growth Prediction using a 3D Morphable Face Model, 2007.
Andreas Steinel: HDR Acquisition, Processing, and Tone Mapping of Astronomical Images, 2009.

Dorotea Dudaš: Vortex Core Line Extraction of 3D Vector Fields Based on Minimal Binding Energy, 2008.
Jens Kerber: Digital Art of Bas-Relief Sculpting, 2008.
Miguel Granados: Background estimation from photographs with application to ghost removal in high dynamic range image reconstruction, 2008.
Henning Peters: Hardware and Software Extensions for a FTIR Multi-Touch Interface, 2008.
Michael Heinz: Descattering in Confocal Imaging, 2009.

Bachelor Theses

Michael Arnold: GPU-based 3D Light Marker Tracking, 2007.
Manuel Schilz: GPU Multi-Marker Tracking using Hierarchical Feature Clustering, 2008.
Alexander Reuter: Design and Implementation of an Efficient Scientific HDR Image Viewer, 2009.

Project Classes:

Towards Automatically Learning 2D shape orientation (Danyi Wang)

30.15 Dissertations, Offers, Awards

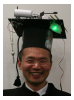
30.15.1 Dissertations

Completed and Defended

Lukas Ahrenberg: Methods for Transform, Analysis and Rendering of Complete Light Representations, 16.07.2007.

Thomas Annen: Efficient Shadow Map Filtering, 12.12.2008.

Robert Bargmann: Learning-Based Speech Animation, 28.11.2008.



Tongbo Chen: New 3D Scanning Techniques for Complex Scenes, 02.12.2008.

Edilson de Aguiar: Animation and Performance Capture Using Digitized Models, 22.12.2008.

Martin Fuchs: Advanced Methods for Relightable Scene Representations in Image Space, 15.12.2008.

Ivo Ihrke: Reconstruction and Rendering of Time-Varying Natural Phenomena, 21.05.2007.

Grzegorz Krawczyk: Perception- inspired Tone Mapping, 30.11.2007.

Torsten Langer: On Generalized Barycentric Coordinates and Their Applications in Geometric Modeling, 18.12.2008.

Andrei Lintu: Realistic Rendering and Reconstruction of Astronomical Objects and an Augmented Reality Application for Astronomy, 07.12.2007.

Volker Scholz: New Editing Techniques for Video Post-Processing, 21.05.2007.

Kuangyu Shi: Path-Line Oriented Visualization of Dynamical Flow Fields, 10.12.2008.

Kaleigh Smith: Contours and Contrast, 11.12.2008.

Wolfram von Funck: Shape Deformations Based on Vector Fields, 22.12.2008.

Akiko Yoshida: Evaluation and Enchancement of HDR Image Appearance on Displays of Varing Dynamic Range, 16.12.2008.

Rhaleb Zayer: Numerical and Variational Aspects of Mesh Parameterization and Deformation, 17.09.2007.

Completed but not yet Defended

- Naveed Ahmed: High Quality Dynamic Reflectance and Surface Reconstruction from Video.
- Christian Fuchs: Capturing and Reconstructing the Appearance of Complex 3D Scenes.
- Jürgen Gall: Filtering and Optimization Strategies for Markerless Human Motion Capture with Skeleton-based Shape Models.
- Oliver Schall: Robust and Efficient Processing Techniques for Static and Dynamic Geometric Data.

30.15.2 Offers for Faculty Positions

- Alexander Belyaev, Reader, Heriot-Watt-University, Edinburgh, UK, 2008
- Hongbo Fu, Assistant Professor, City University of Hong Kong, 2009
- Michael Goesele, W1-Professorhip, TH Darmstadt, 2007
- Thorsten Grosch, W1-Professorship, Univ. Magdeburg, 2009
- Hendrik Lensch, W3-Professorship, Univ. Ulm, 2009
- Bodo Rosenhahn, W3-Professorship, Univ. Hannover, 2008
- Rhaleb Zayer, Researcher, LORIA/INRIA Lorraine, France, 2008

30.15.3 Awards

- Edilson de Aguiar, Disney/CMU Postdoc Fellowship, 2009
- Tongbo Chen, USC Postdoc Fellowship, 2008
- Zhao Dong, Best Ranked Paper, Computer Graphics International, 2008
- Zhao Dong, Chinese Government Award for Outstanding Self-Financed Students Abroad, 2008
- Jürgen Gall, DAGM Main Prize, 2007
- Jürgen Gall, ETH Postdoc Fellowship, 2009
- Michael Goesele, Eurographics Young Researcher Award 2008
- Ivo Ihrke, Humboldt Postdoc Fellowship, 2007
- Ivo Ihrke, Eduard Martin Prize, 2007
- Jens Kerber, 2nd Best Paper Award, SCCG, 2007
- Grzegorz Krawczyk, 2nd Best Paper Award, Eurographics, 2007
- Hendrik Lensch, DFG Emmy-Nother-Fellowship, 2007
- Rafal Mantiuk, UBC Postdoc Fellowship, 2008
- Bodo Rosenhahn, DAGM Olympus-Prize, 2007
- Oliver Schall, Best Paper Award, ACM Symp. Solid and Physical Modeling, 2007
- Thomas Schulz, Best Paper Award, IEEE Visualization, 2008
- Thomas Schulz, DAAD Postdoc Fellowship, 2009
- Kaleigh Smith, 3rd Best Paper Award, Eurographics, 2008
- Christian Theobalt, Eurographics Young Researcher Award 2009
- Thorsten Thormählen, Group of Eight Australia/Germany Joint Research Cooperation Scheme, 2009

30.16 Grants and Cooperations

- Alexander Belyaev, Hans-Peter Seidel, Principal Investigators, EU NoE "Advanced and Innovative Models And Tools for the development of Semantic-based systems for Handling, Acquiring, and Processing knowledge Embedded in multidimensional digital objects (AIM@SHAPE)", funded by EU

- Elmar Eisemann, Junior Research Group ECLEXIS, Cluster of Excellence on Multimodal Computing and Interaction, funded by DFG
- Martin Fuchs, Passive Reflectance field Displays, industrial collaboration with Mitsubishi Research Electric Laboratories (MERL)
- Thorsten Grosch, RayLight: Hybrid Rendering with Real-time Ray Tracing and Global Illumination, industrial collaboration with Samsung and DFKI
- Hendrik Lensch, Junior Research Group on Acquisition, Modification and Rendering of Reflectance Fields, funded by DFG (Emmy Noether Program)
- Meinard Müller, Junior Research Group Multimedia Information Retrieval and Music Processing, Cluster of Excellence on Multimodal Computing and Interaction, funded by DFG
- Meinard Müller, Reconstruction of Motion Sequences (Rekonstruktion von Bewegungsabläufen, REKOBA), funded by DFG
- Meinard Müller, Automated Processing of Music Documents (Automatisierte Erschließung von Musikdokumenten, ARMADA), funded by DFG
- Meinard Müller, Stabilized Motion Analysis through Retrieval Techniques (SMART), funded by DFG
- Karol Myszkowski, Backward Compatible High Dynamic Range Video Compression, industrial collaboration with BrightSide Technologies (Dolby)
- Bodo Rosenhahn, Jürgen Gall, Crash Test Video Analysis, industrial collaboration with Daimler
- Bodo Rosenhahn, Junior Research Group on Markerless Motion Capture, Max Planck Center for Visual Computing and Communication, funded by DFG
- Hans-Peter Seidel, Principal Investigator, EU NoE "3DTV-Integrated Three-Dimensional Television - Capture, Transmission and Display (3DTV)", funded by EU
- Hans-Peter Seidel, Scientific Coordinator, Cluster of Excellence on Multimodal Computing and Interaction, funded by DFG
- Hans-Peter Seidel, Principal Investigator, Cluster of Excellence on Multimodal Computing and Interaction, funded by DFG
- Hans-Peter Seidel, Co-Director, Max Planck Center for Visual Computing and Communication, funded by BMBF, MPG and Stanford University
- Michael Sips, Junior Research Group on Explorative Data Analysis, Max Planck Center for Visual Computing and Communication, funded by BMBF
- Robert Strzodka, Junior Research Group on Integrative Scientific Computing, Max Planck Center for Visual Computing and Communication, funded by BMBF

- Thorsten Thormählen, Junior Research Group on Image-based 3D Scene Analysis, Max Planck Center for Visual Computing and Communication, funded by BMBF
- Thorsten Thormählen, Real Time Stereo System for Face Detection and Tracking, industrial collaboration with Microsoft Research
- Michael Wand, Junior Research Group on Statistical Geometry Processing, Cluster of Excellence on Multimodal Computing and Interaction, funded by DFG

30.17 Publications

Books

[1] A. G. Belyaev and M. Garland, eds. *Proceedings of the 5th Eurographics Symposium on Geometry Processing*, Aire-la-Ville, Switzerland, 2007, *ACM International Conference Proceeding Series*, vol. 257. Eurographics Association.

[2] S. Creem-Regehr and K. Myszkowski, eds. *Symposium on Applied Perception in Graphics and Visualization*. ACM, New York, USA, 2008.

[3] A. Elgammal, B. Rosenhahn, and R. Klette, eds. *2nd Workshop: Human Motion - Understanding, Modeling, Capture and Animation*, *LNCS 4814*. Springer, Berlin, Germany, 2007.

[4] H. P. A. Lensch, B. Rosenhahn, H.-P. Seidel, P. Slusallek, and J. Weickert, eds. *Vision, Modeling, and Visualization*. Die Deutsche Bibliothek, Max-Planck-Gesellschaft, 2007.

[5] K. Myszkowski, ed. *Proceedings of the 24th Spring Conference on Computer Graphics (SCCG '08)*. ACM, New York, USA, 2008.

[6] K. Myszkowski, R. Mantiuk, and G. Krawczyk. *High Dynamic Range Video*. Synthesis Digital Library of Engineering and Computer Science. Morgan & Claypool Publishers, San Rafael, USA, 2008.

[7] B. Rosenhahn, R. Klette, and D. Metaxas, eds. *Human Motion - Understanding, Modeling, Capture and Animation*, *Computational Imaging and Vision*, vol. 36. Springer, Dordrecht, The Netherlands, 2008.

Journal articles and book chapters

[1] E. de Aguiar, C. Stoll, N. Ahmed, and H.-P. Seidel. Performance capture from sparse multi-view video. In G. Turk, ed., *Proceedings of ACM SIGGRAPH 2008*, Los Angeles, USA, 2008, *ACM Transactions on Graphics*, vol. 27(3), pp. 1–10. ACM.

[2] T. Annen, Z. Dong, T. Mertens, P. Bekaert, H.-P. Seidel, and J. Kautz. Real-time, all-frequency shadows in dynamic scenes. In G. Turk, ed., *Proceedings of ACM SIGGRAPH 2008*, Los Angeles, USA, 2008, *ACM Transactions on Graphics*, vol. 27, pp. 1–8. ACM.

[3] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Dynamic range independent image quality assessment. In G. Turk, ed., *Proceedings of ACM SIGGRAPH 2008*, Los Angeles, USA, 2008, *ACM Transactions on Graphics*, vol. 27(3), pp. 1–10. ACM.

[4] T. O. Aydin, K. Myszkowski, and H.-P. Seidel. Predicting display visibility under dynamically changing lighting conditions. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28(2):173–182, 2009.

[5] M. Bokeloh, A. Berner, M. Wand, H.-P. Seidel, and A. Schilling. Symmetry detection using line features. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28(2):697–706, 2009.

[6] T. Brox, B. Rosenhahn, and D. Cremers. Contours, optic flow, and prior knowledge: Cues for capturing 3D human motion in videos. In B. Rosenhahn, R. Klette, and D. Metaxas, eds., *Human Motion - Modeling, Tracking, Capture and Animation*, *Computational Imaging and Vision*, vol. 36, ch. 11, pp. 265–293. Springer, Dordrecht, The Netherlands, 2008.

[7] N. Cuntz, A. Kolb, R. Strzodka, and D. Weiskopf. Particle level set advection for the interactive visualization of unsteady 3D flow. *Computer Graphics Forum (Proc. EuroVis)*, 27(3):719–726, 2008.

[8] P. Didyk, R. Mantiuk, M. Hein, and H.-P. Seidel. Enhancement of bright video features for hdr displays. *Computer Graphics Forum*, 27(4):1265–1274, 2008.

[9] C. Dyken, G. Ziegler, C. Theobalt, and H.-P. Seidel. High-speed marching cubes using histopyramids. *Computer Graphics Forum*, 27(8):2028–2039, 2008.

[10] M. Eisemann, B. de Decker, M. A. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt, and A. Sellent. Floating textures. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 27(2):409–418, 2008.

[11] C. Fuchs, T. Chen, M. Goesele, H. Theisel, and H.-P. Seidel. Density estimation for dynamic volumes. *Computers and Graphics*, 31(2):205–211, 2007.

[12] C. Fuchs, M. Heinz, M. Levoy, H.-P. Seidel, and H. P. A. Lensch. Combining confocal imaging and descattering. *Computer Graphics Forum (Proc. Eurographics Symposium on Rendering)*, 27(4):1245–1253, 2008.

[13] M. Fuchs, V. Blanz, H. P. A. Lensch, and H.-P. Seidel. Adaptive sampling of reflectance fields. *ACM Transactions on Graphics (TOG)*, 26(2):1–18, 2007.

[14] M. Fuchs, H. P. A. Lensch, V. Blanz, and H.-P. Seidel. Superresolution reflectance fields: Synthesizing images for intermediate light directions. In D. Cohen-Or and P. Slavík, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Prague, Czech Republic, 2007, vol. 26(3), pp. 447–456. Blackwell.

[15] M. Fuchs, R. Raskar, H.-P. Seidel, and H. P. A. Lensch. Towards passive 6D reflectance field displays. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 27(3):58, 2008.

[16] W. von Funck, T. Weinkauf, H. Theisel, and H.-P. Seidel. Smoke surfaces: An interactive flow visualization technique inspired by real-world flow experiments. *IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE Visualization)*, 14(6):1396–1403, 2008.

[17] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture - a multi-layer framework. *International Journal of Computer Vision*, pp. 1–18, 2009. To appear.

[18] S. Gehrig, B. Hernán, and J. Gall. Accurate and model-free pose estimation of crash test dummies. In R. Klette, D. Metaxas, and B. Rosenhahn, eds., *Human Motion - Understanding, Modeling, Capture and Animation*, pp. 453–473. Springer, Heidelberg, 2008.

[19] D. Göddeke, R. Strzodka, J. Mohd-Yusof, P. McCormick, S. H. Buijssen, M. Grajewski, and S. Turek. Exploring weak scalability for FEM calculations on a GPU-enhanced cluster. *Parallel Computing*, 33(10-11):685–699, 2007.

[20] D. Göddeke, R. Strzodka, J. Mohd-Yusof, P. McCormick, H. Wobker, C. Becker, and S. Turek. Using GPUs to improve multigrid solver performance on a cluster. *International Journal of Computational Science and Engineering*, 4(1):36–55, 2008.

[21] D. Göddeke, R. Strzodka, and S. Turek. Performance and accuracy of hardware-oriented native-, emulated- and mixed-precision solvers in FEM simulations. *International Journal of Parallel, Emergent and Distributed Systems*, 22(4):221–256, 2007.

[22] M. Goesele and K. Myszkowski. Hdr applications in computer graphics. In B. Hoefflinger, ed., *High-Dynamic-Range (HDR) Vision*, Springer Series in Advanced Microelectronics (26), pp. 193–210. Springer, Heidelberg, 2007.

[23] B. Goldluecke, I. Ihrke, C. Linz, and M. Magnor. Weighted minimal hypersurface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1194–1208, 2007.

[24] S. Hahmann, A. Belyaev, L. Buse, G. Elber, B. Mourrain, and C. Rössl. Shape interrogation. In L. de Floriani and M. Spagnuolo, eds., *Shape Analysis and Structuring*, Mathematics and Visualization, ch. 1, pp. 1–52. Springer, Berlin, Germany, 2008.

[25] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Munich, Germany, 2009, vol. 2, pp. 337–346. Blackwell.

[26] R. Herzog, V. Havran, S. Kinuwaki, K. Myszkowski, and H.-P. Seidel. Global illumination using photon ray splatting. In D. Cohen-Or and P. Slavik, eds., *The European Association for Computer Graphics 28th Annual Conference: EUROGRAPHICS 2007*, Prague, Czech Republic, 2007, *Computer Graphics Forum*, vol. 26(3), pp. 503–513. Blackwell.

[27] R. Herzog, S. Kinuwaki, K. Myszkowski, and H.-P. Seidel. Render2MPEG: a perception-based framework towards integrating rendering and video compression. In R. Scopigno and E. Gröller, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Crete, Greece, 2008, vol. 27(2), pp. 183–192. Blackwell.

[28] R. Herzog, K. Myszkowski, and H.-P. Seidel. Anisotropic radiance-cache splatting for efficiently computing high-quality global illumination with lightcuts. In M. Stamminger and P. Dutré, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, München, Germany, 2009, vol. 28, pp. 259–268. Wiley-Blackwell.

[29] M. B. Hullin, M. Fuchs, I. Ihrke, H.-P. Seidel, and H. P. A. Lensch. Fluorescent immersion range scanning. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2008)*, Los Angeles, USA, 2008, vol. 27, pp. 87:1–87:10. ACM.

[30] I. Ihrke, G. Ziegler, A. Tevs, C. Theobalt, M. Magnor, and H.-P. Seidel. Eikonal rendering: Efficient light transport in refractive objects. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 26(3):59–1 – 59–8, 2007.

[31] B. Kim, K. H. Lee, K. J. Kim, R. Mantiuk, S. Hahn, T. J. Kim, and Y. H. Kim. Prediction of perceptible artifacts in jpeg 2000-compressed chest ct images using mathematical and perceptual quality metrics. *American Journal of Roentgenology*, 190:328–334, 2008.

[32] B. Kim, K. H. Lee, K. J. Kim, R. Mantiuk, H.-r. Kim, and Y. H. Kim. Artifacts in slab average-intensity-projection images reformatted from jpeg 2000 compressed thin-section abdominal ct data sets. *American Journal of Roentgenology*, 190:W342–W350, 2008.

[33] K. J. Kim, B. Kim, K. H. Lee, T. J. Kim, R. Mantiuk, H.-S. Kang, and Y. H. Kim. Regional difference in compression artifacts in low-dose chest ct images: Effects of mathematical and perceptual factors. *American Journal of Roentgenology*, 191:W30–W37, 2008.

[34] G. Krawczyk, K. Myszkowski, and D. Brosch. Hdr tone mapping. In B. Hoefflinger, ed., *High-Dynamic-Range (HDR) Vision*, Springer Series in Advanced Microelectronics (26), pp. 147–178. Springer, Heidelberg, 2007.

[35] G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Contrast restoration by adaptive countershading. In D. Cohen-Or and P. Slavik, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Prague, Czech Republic, 2007, vol. 26(3), pp. 581–590. Blackwell.

[36] B. Krüger, J. Tautges, M. Müller, and A. Weber. Multi-mode tensor representation of motion data. *Journal of Virtual Reality and Broadcasting*, 5(5):1–13, 2008.

[37] F. Kurth and M. Müller. Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, 2008.

[38] T. Langer and H.-P. Seidel. Higher order barycentric coordinates. In G. Drettakis and R. Scopigno, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Crete, Greece, 2008, vol. 27(2), pp. 459–466. Blackwell.

[39] X. Liu, Z. Dong, H. Bao, and Q. Peng. Caustic spot light for rendering caustics. *The Visual Computer*, 24(7-9):485–494, 2008.

[40] R. Mantiuk. Hdr image and video compression. In B. Hoefflinger, ed., *High-Dynamic-Range (HDR) Vision*, Springer Series in Advanced Microelectronics (26), pp. 179–192. Springer, Heidelberg, 2007.

[41] R. Mantiuk, S. Daly, and L. Kerofsky. Display adaptive tone mapping. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 27(3):68, 2008.

[42] R. Mantiuk, R. Mantiuk, A. Tomaszewska, and W. Heidrich. Color correction for tone mapping. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28(3), 2009.

[43] R. Mantiuk and H.-P. Seidel. Modeling a generic tone-mapping operator. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 27(2):699–708, 2008.

[44] M. Müller and M. Jimbo. Cyclic sequences of k-subsets with distinct consecutive unions. *Discrete Mathematics*, 308(2-3):457–464, 2008.

[45] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics*, 27(1):5, 2008.

[46] M. Okabe, K. Anjyo, T. Igarashi, and H.-P. Seidel. Animating pictures of fluid using video examples. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28, 2009.

[47] S. Popov, J. Günther, H.-P. Seidel, and P. Slusallek. Stackless kd-tree traversal for high performance GPU ray tracing. In D. Cohen-Or and P. Slavik, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Prague, Czech Republic, 2007, vol. 26(3), pp. 415–424. Blackwell.

[48] T. Ritschel, M. Ihrke, J. R. Frisvad, J. Coppens, K. Myszkowski, and H.-P. Seidel. Temporal glare: Real-time dynamic simulation of the scattering in the human eye. *Computer Graphics Forum (Proc. EUROGRAPHICS 2009)*, 28(3):183–192, 2009.

[49] T. Ritschel, K. Smith, M. Ihrke, T. Grosch, K. Myszkowski, and H.-P. Seidel. 3d unsharp masking for scene coherent enhancement. In G. Turk, ed., *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, Los Angeles, USA, 2008, vol. 27(3), pp. 1–8. ACM.

[50] B. Rosenhahn, T. Brox, and J. Weickert. Three-dimensional shape knowledge for joint image segmentation and pose tracking. *International Journal of Computer Vision*, 73(3):243–262, 2007.

[51] B. Rosenhahn, U. Kersting, K. Powell, T. Brox, and H.-P. Seidel. Tracking clothed people. In B. Rosenhahn, R. Klette, and D. Metaxas, eds., *Human Motion Understanding,. Modeling, Capture and Animation*, *Computational Imaging and Vision*, vol. 36, ch. 12, pp. 295–317. Springer, Dordrecht, The Netherlands, 2008.

[52] O. Schall, A. G. Belyaev, and H.-P. Seidel. Adaptive feature-preserving non-local denoising of static and time-varying range data. *Computer-Aided Design*, 40(1):701–707, 2008.

[53] K. Scherbaum, M. Sunkel, H.-P. Seidel, and V. Blanz. Prediction of individual non-linear aging trajectories of faces. In D. Cohen-Or and P. Slavik, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Prague, Czech Republic, 2007, vol. 26(3), pp. 285–294. Blackwell.

[54] T. Schultz, N. Sauber, A. Anwander, H. Theisel, and H.-P. Seidel. Virtual klingler dissection: Putting fibers into context. *Computer Graphics Forum (Proc. EuroVis)*, 27(3):1063–1070, 2008.

[55] T. Schultz and H.-P. Seidel. Estimating crossing fibers: A tensor decomposition approach. *IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE Visualization)*, 14(6):1635–1642, 2008.

[56] T. Schultz, H. Theisel, and H.-P. Seidel. Topological visualization of brain diffusion MRI data. *IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE Visualization)*, 13(6):1496–1503, 2007.

[57] T. Schultz, J. Weickert, and H.-P. Seidel. A higher-order structure tensor. In D. H. Laidlaw and J. Weickert, eds., *Visualization and Processing of Tensor Fields - Advances and Perspectives*, Mathematics and Visualization, pp. 263–280. Springer, Berlin, 2009.

[58] H.-P. Seidel. Excellence cluster "multimodal computing and interaction". *IT Information Technology*, 50(04):253–257, 2008.

[59] K. Shi, H. Theisel, H. Hauser, T. Weinkauf, K. Matkovic, H.-C. Hege, and H.-P. Seidel. Path line attributes - an information visualization approach to analyzing the dynamic behavior of 3D time-dependent flow fields. In H.-C. Hege, K. Polthier, and G. Scheuermann, eds., *Topology-Based Methods in Visualization II*, Grimma, Germany, 2009, Springer series of Mathematics and Visualization, pp. 75–88. Springer.

[60] K. Smith, P.-E. Landes, J. Thollot, and K. Myszkowski. Apparent greyscale: A simple and fast conversion to perceptually accurate images and video. In R. Scopigno and E. Gröller, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Crete, Greece, 2008, vol. 27(2), pp. 193–200. Blackwell.

[61] M. Song, Z. Dong, C. Theobalt, H. Wang, Z. Liu, and H.-P. Seidel. A general framework for efficient 2d and 3d facial expression analogy. *IEEE Transactions on Multimedia*, 9(7):1384–1395, 2007.

[62] E. Stoykova, A. A. Alatan, P. Benzie, N. Grammalidis, S. Malassiotis, J. Ostermann, S. Piekh, V. Sainov, C. Theobalt, T. Thevar, and X. Zabulis. 3-d time-varying scene capture technologies - a survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11):1568–1586, 2007.

[63] H. Theisel, T. Weinkauf, H.-C. Hege, and H.-P. Seidel. On the applicability of topological methods for complex flow data. In H. Hauser, H. Hagen, and H. Theisel, eds., *Topology-based methods in visualization*, Mathematics and Visualization, pp. 105–120. Springer, Berlin, Germany, 2007.

[64] C. Theobalt, E. de Aguiar, M. Magnor, and H.-P. Seidel. Reconstructing human shape, motion and appearance from multi-view video. In H. Ozaktas and L. Onural, eds., *Three-Dimensional Television: Capture, Transmission, and Display*, pp. 29–58. Springer, Heidelberg, Germany, 2008.

[65] C. Theobalt, N. Ahmed, G. Ziegler, and H.-P. Seidel. High-quality reconstruction of virtual actors from multi-view video streams. *IEEE Signal Processing Magazine*, 24(6):45–57, 2007.

[66] T. Thormählen and H.-P. Seidel. 3d-modeling by ortho-image generation from image sequences. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 27(3):86:1–86:5, 2008.

[67] M. Wand. Rendering of very large models. In M. Gross and H. Pfister, eds., *Point-Based Graphics*, pp. 313–326. Morgan Kaufmann/Elsevier, Amsterdam, The Netherlands, 2007.

[68] M. Wand, A. Berner, M. Bokeloh, P. Jenke, A. Fleck, M. Hoffmann, B. Maier, D. Staneker, A. Schilling, and H.-P. Seidel. Processing and interactive editing of huge point clouds from 3d scanners. *Computers and Graphics*, 32(2):204–220, 2008.

[69] T. Weinkauf, H. Theisel, H.-C. Hege, and H.-P. Seidel. Feature flow fields in out-of-core settings. In H. Hauser, H. Hagen, and H. Theisel, eds., *Topology-based methods in visualization*, Mathematics and Visualization, pp. 51–63. Springer, Berlin, Germany, 2007.

[70] S. Yoshizawa, A. Belyaev, and H.-P. Seidel. Skeleton-based variational mesh deformations. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 26(3):255–264, 2007.

[71] S. Yoshizawa, A. Belyaev, H. Yokota, and H.-P. Seidel. Fast, robust, and faithful methods for detecting crest lines on meshes. *Computer Aided Geometric Design*, 25(8):545–560, 2008.

[72] R. Ziegler, S. Bucheli, L. Ahrenberg, M. Magnor, and M. Gross. A bidirectional light field - hologram transform. In D. Cohen-Or and P. Slavik, eds., *Computer Graphics Forum (Proc. EUROGRAPHICS)*, Prague, Czech Republic, 2007, vol. 26(3), pp. 435–446. Blackwell.

Conference articles

[1] B. Adams, M. Ovsjanikov, M. Wand, H.-P. Seidel, and L. Guibas. Meshless modeling of deformable shapes and their motion. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Dublin, Ireland, 2008, pp. 77–86. Eurographics Association.

[2] E. de Aguiar, R. Zayer, C. Theobalt, M. Magnor, and H.-P. Seidel. A simple framework for natural animation of digitized models. In *SIBGRAPI'07 - XX Brazilian Symposium on Computer Graphics and Image Processing*, Belo Horizonte, Brazil, 2007, pp. 3–10. IEEE.

[3] N. Ahmed, C. Theobalt, P. Dobrev, H.-P. Seidel, and S. Thrun. Robust fusion of dynamic shape and normal capture for high-quality reconstruction of time-varying geometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage, Alaska, 2008, pp. 1–8. IEEE Computer Society.

[4] N. Ahmed, C. Theobalt, M. Magnor, and H.-P. Seidel. Spatio-temporal registration techniques for relightable 3d video. In *International Conference on Image Processing 2007*, San Antonio, TX, USA, 2007, vol. 2, pp. 501–504. IEEE.

[5] N. Ahmed, C. Theobalt, C. Rössl, S. Thrun, and H.-P. Seidel. Dense correspondence finding for parametrization-free animation reconstruction from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage, Alaska, 2008, pp. 1–8. IEEE Computer Society.

[6] B. Ajdin, M. B. Hullin, C. Fuchs, H.-P. Seidel, and H. P. A. Lensch. Demosaicing by smoothing along 1d features. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

[7] T. Annen, T. Mertens, P. Bekaert, H.-P. Seidel, and J. Kautz. Convolution shadow maps. In J. Kautz and S. Pattanaik, eds., *Rendering Techniques 2007: Eurographics Symposium on Rendering*, Grenoble, France, 2007, *Eurographics / ACM SIGGRAPH Symposium Proceedings*, vol. 18, pp. 51–60. Eurographics.

[8] T. Annen, T. Mertens, H.-P. Seidel, E. Flerackers, and J. Kautz. Exponential shadow maps. In C. Shaw and L. Bartram, eds., *Proceedings of Graphics Interface 2008*, Windsor, Ontario, Canada, 2008, *ACM International Conference Proceeding Series*, vol. 34, pp. 155–161. A K Peters.

[9] T. Annen, H. Theisel, C. Rössl, G. Ziegler, and H.-P. Seidel. Vector field contours. In C. Shaw and L. Bartram, eds., *Proceedings of the Graphics Interface 2008*, Windsor, Ontario, Canada, 2008, pp. 97–105. ACM.

[10] B. Atcheson, I. Ihrke, W. Heidrich, A. Tevs, D. Bradley, M. Magnor, and H.-P. Seidel. Time-resolved 3d capture of non-stationary gas flows. In *SIGGRAPH Asia '08: ACM SIGGRAPH Asia 2008 papers*, Singapore, 2008, pp. 1–9. ACM.

[11] T. O. Aydin, R. Mantiuk, and H.-P. Seidel. Extending quality metrics to full luminance range images. In A. Heimer, ed., *Human Vision and Electronic Imaging XIII*, San Jose, USA, 2008. SPIE.

[12] A. Baak, M. Müller, and H.-P. Seidel. An efficient algorithm for keyframe-based motion retrieval in the presence of temporal deformations. In M. S. Lew, A. Del Bimbo, and E. M. Bakker, eds., *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval*, Vancouver, British Columbia, Canada, 2008, pp. 451–458. ACM.

[13] F. Banterle, K. Debattista, A. Artusi, S. Pattanaik, K. Myszkowski, P. Ledda, M. Bloj, and A. Chalmers. High dynamic range imaging and ldr expansion for generating hdr content. In *EUROGRAPHICS State-of-the-Art Report*, Munich, 2009, pp. 17–44. Eurographics.

[14] A. Berner, M. Bokeloh, M. Wand, A. Schilling, and H.-P. Seidel. A graph-based approach to symmetry detection. In *Symposium on Volume and Point-Based Graphics*, Los Angeles, CA, 2008, pp. 1–8. Eurographics Association.

[15] V. Blanz, K. Scherbaum, and H.-P. Seidel. Fitting a morphable model to 3d scans of faces. In *Eleventh IEEE International Conference on Computer Vision*, Rio de Janeiro, Brasil, 2007, *DVD Proceedings*, vol. CFP07198-CDR, pp. 1–8. IEEE ICCV 2007, Omnipress.

[16] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. Nonparametric density estimation with adaptive anisotropic kernels for human motion tracking. In A. Elgammal, B. Rosenhahn, and R. Klette, eds., *2nd Workshop on Human Motion*, Rio de Janeiro, Brazil, 2007, *LNCS 4814*, pp. 152–165. Springer.

[17] T. Chen, H.-P. Seidel, and H. P. A. Lensch. Modulated phase-shifting for 3D scanning. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage, Alaska, USA, 2008. IEEE Computer Society.

[18] C. Crassin, F. Neyret, S. Lefebvre, and E. Eisemann. Gigavoxels: Ray-guided streaming for efficient and detailed voxel rendering. In *ACM Symposium on Interactive 3D Graphics and Games (i3D)*, Boston, USA, 2009, pp. 15–22. ACM.

[19] D. Damm, C. Fremerey, F. Kurth, M. Müller, and M. Clausen. Multimodal presentation and browsing of music. In *IMCI '08: Proceedings of the 10th International Conference on Multimodal Interfaces*, Chania, Crete, Greece, 2008, pp. 205–208. ACM.

[20] D. Damm, C. Fremerey, F. Kurth, M. Müller, and M. Clausen. SyncPlayer - Multimodale Wiedergabe, Navigation und Suche in heterogenen digitalen Musikkollektionen. In T. Mandl, N. Fuhr, and A. Henrich, eds., *Proceedings of the Workshop on Lernen-Wissen-Adaptivität (LWA 2008)*, Würzburg, Germany, 2008, pp. 1–8. GI.

[21] E. Danovaro, L. Papaleo, D. Sobrero, M. Attene, and W. Saleem. Advanced remote inspection and download of 3D shapes. In *Proceedings of the twelfth international conference on 3D web technology (Web3D '07:)*, Perugia, Italy, 2007, pp. 57–60. ACM.

[22] Z. Dong, J. Kautz, C. Theobalt, and H.-P. Seidel. Interactive global illumination using implicit visibility. In M. Alexa, T. Ju, and S. Gortler, eds., *The 15th Pacific Conference on Computer Graphics and Applications (Pacific Graphics 2007)*, Maui Hawaii, USA, 2007, pp. 44–54. IEEE Computer Society.

[23] S. Ewert and M. Müller. Refinement strategies for music synchronization. In *Proceedings of the 5th International Symposium on Computer Music Modeling and Retrieval (CMMR 2008)*, Copenhagen, Denmark, 2008, Lecture Notes in Computer Science. Springer. To appear.

[24] S. Ewert, M. Müller, and P. Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009. IEEE.

[25] C. Fremerey, M. Müller, F. Kurth, and S. Ewert. Automatic mapping of scanned sheet music to audio recordings. In J. P. Bello, E. Chew, and D. Turnbull, eds., *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, USA, 2008, pp. 413–418. ISMIR.

[26] W. von Funck, H. Theisel, and H.-P. Seidel. Elastic secondary deformations by vector field integration. In A. Belyaev and M. Garland, eds., *Proceedings of the 5th Eurographics Symposium on Geometry Processing*, Barcelona, Spain, 2007, *ACM International Conference Proceeding Series*, vol. 257, pp. 99–108. ACM.

[27] W. von Funck, H. Theisel, and H.-P. Seidel. Explicit control of vector field based shape deformations. In M. Alexa, S. Gortler, and T. Ju, eds., *Proceedings of the Pacific Conference on Computer Graphics and Applications, Pacific Graphics 2007*, Maui, Hawaii, 2007, pp. 291–300. IEEE Computer Society.

[28] W. von Funck, H. Theisel, and H.-P. Seidel. Implicit boundary control of vector field based shape deformations. In R. Martin, M. Sabin, and J. Winkler, eds., *Mathematics of Surfaces XII, 12th IMA International Conference*, Sheffield, UK, 2007, *LNCS 4647*, pp. 154–165. Springer.

[29] W. von Funck, H. Theisel, and H.-P. Seidel. Volume-preserving mesh skinning. In O. Deussen, D. Keim, and D. Saupe, eds., *13th International Fall Workshop on Vision, Modeling and Visualization*, Konstanz, Germany, 2008, pp. 407–414. Akademische Verlagsgesellschaft AKA.

[30] J. Gall, B. Rosenhahn, S. Gehrig, and H.-P. Seidel. Model-based motion capture for crash test video analysis. In G. Rigoll, ed., *Pattern Recognition*, Munich, Germany, 2008, *LNCS 5096*, pp. 92–101. Springer.

[31] J. Gall, B. Rosenhahn, and H.-P. Seidel. Drift-free tracking of rigid and articulated objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, USA, 2008, pp. 1–8. IEEE Computer Society.

[32] M. Goesele. Images, images, billions of images. In H. P. Lensch, B. Rosenhahn, H.-P. Seidel, P. Slusallek, and J. Weickert, eds., *Proceedings of the Vision, Modeling, and Visualization Conference 2007 (VMV 2007)*, Saarbrücken, Germany, 2007, pp. 117–118. Aka GmbH.

[33] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *Proceedings of the 11th International Conference on Computer Vision (ICCV 2007)*, Rio de Janeiro, Brazil, 2007, pp. 265–270. IEEE.

[34] M. A. Granados V., H.-P. Seidel, and H. P. A. Lensch. Graphics interface. In C. Shaw and L. Bartram, eds., *Proceedings of the Graphics Interface 2008 Conference*, Windsor, Ontario, Canada, 2008, ACM International Conference Proceeding Series, pp. 33–40. ACM Press.

[35] T. Grosch, T. Eble, and S. Mueller. Consistent interactive augmentation of live camera images with correct near-field illumination. In S. N. Spencer, ed., *VRST 2007 (ACM Symposium on Virtual Reality Software and Technology)*, Newport Beach, California, USA, 2007, pp. 125–132. ACM.

[36] J. Günther, S. Popov, H.-P. Seidel, and P. Slusallek. Realtime ray tracing on gpu with bvh-based packet traversal. In A. Keller and P. Christensen, eds., *IEEE/Eurographics Symposium on Interactive Ray Tracing 2007*, Ulm, Germany, 2007, pp. 113–118. IEEE.

[37] N. Hasler and K.-P. Hasler. Short-term tide prediction. In F. Hamprecht, C. Schnörr, and B. Jähne, eds., *Pattern Recognition, 29th DAGM Symposium*, Heidelberg, Germany, 2007, *LNCS 4713*, pp. 375–384. Springer.

[38] R. Herzog and H.-P. Seidel. Lighting details preserving photon density estimation. In M. Alexa, T. Ju, and S. Gortler, eds., *The 15th Pacific Conference on Computer Graphics and Application*, Maui, Hawaii, USA, 2007, pp. 407–410. IEEE Computer Society.

[39] M. B. Hullin, M. Fuchs, I. Ihrke, B. Ajdin, H.-P. Seidel, and H. P. A. Lensch. Direct visualization of real-world light transport. In O. Deussen, D. Keim, and D. Saupe, eds., *13th International Fall Workshop on Vision, Modeling and Visualization*, Konstanz, Germany, 2008, pp. 363–371. Akademische Verlagsgesellschaft AKA.

[40] M. Ihrke, T. Ritschel, K. Smith, T. Grosch, K. Myszkowski, and H.-P. Seidel. A perceptual evaluation of 3d unsharp masking. In B. E. Rogowitz and T. N. Pappas, eds., *Human Vision and Electronic Imaging XIV, IS&T/SPIE's 21st Annual Symposium on Electronic Imaging*, San Jose, USA, 2009, *Annual Symposium on Electronic Imaging*, vol. 7240. SPIE.

[41] P. Jenke, M. Wand, and W. Straßer. Patch-graph reconstruction for piecewise smooth surfaces. In O. Deussen, D. Keim, and D. Saupe, eds., *Proceedings Vision, Modeling and Visualization (VMV 2008)*, Konstanz, Germany, 2008, pp. 3–12. Akademische Verlagsgesellschaft AKA.

[42] J. Kerber, A. Belyaev, and H.-P. Seidel. Feature preserving depth compression of range images. In M. Sbert, ed., *Proceedings of the 23rd Spring Conference on Computer Graphics*, Budmerice, Slovakia, 2007, pp. 110–114. Comenius University, Slovakia. Winner 2nd best SCCG 2007 paper award.

[43] F. Kurth, D. Damm, C. Fremerey, M. Müller, and M. Clausen. A framework for managing multimodal digitized music collections. In *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, Aarhus, Denmark, 2008, *LNCS 5173*, pp. 334–345. Springer.

[44] T. Langer, A. Belyaev, and H.-P. Seidel. Mean value Bézier maps. In F. Chen and B. Jüttler, eds., *Advances in Geometric Modeling and Processing*, Hangzhou, China, 2008, *LNCS 4975*, pp. 231–243. Springer.

[45] T. Langer and H.-P. Seidel. Mean value Bézier surfaces. In R. Martin, M. Sabin, and J. Winkler, eds., *Mathematics of Surfaces XII*, Sheffield, England, 2007, *LNCS 4647*, pp. 263–274. Springer.

[46] A. Lintu, L. Hoffmann, M. Magnor, H. P. A. Lensch, and H.-P. Seidel. 3d reconstruction of reflection nebulae from a single image. In H. P. A. Lensch, B. Rosenhahn, H.-P. Seidel, P. Slusallek, and J. Weickert, eds., *Vision, Modeling, and Visualization*, Saarbrücken, Germany, 2007, pp. 109–116. MPI Informatik, AKA.

[47] A. Lintu, H. P. A. Lensch, M. Magnor, S. El-Abed, and H.-P. Seidel. 3d reconstruction of emission and absorption in planetary nebulae. In H.-C. Hege, R. Machiraju, T. Möller, and M. Sramek, eds., *IEEE/EG International Symposium on Volume Graphics*, Prague, Czech Republic, 2007, pp. 9–16. A. K. Peters.

[48] A. Lintu, H. P. A. Lensch, M. Magnor, T.-H. Lee, S. El-Abed, and H.-P. Seidel. Multi-wavelength-based method to de-project gas and dust distributions of several planetary nebulae. In R. L. Corradi, A. Manchado, and N. Soker, eds., *Proceedings of Asymmetrical Planetary Nebulae IV*, La Palma, Spain, 2008, pp. 1–6. ADS.

[49] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. High dynamic range image and video compression - fidelity matching human visual performance. In *IEEE International Conference on Image Processing (ICIP 2007)*, San Antonio, TX, USA, 2007, vol. 1, pp. 9–12. IEEE.

[50] R. Mantiuk, D. Zdrojewska, A. Tomaszewska, R. Mantiuk, and K. Myszkowski. Selected problems of high dynamic range video compression and gpu-based contrast domain tone mapping. In K. Myszkowski, ed., *SCCG '08: Proceedings of the 24th Spring Conference on Computer Graphics*, Budmerice, Slovakia, 2008, pp. 11–18. ACM.

[51] M. Müller and D. Appelt. Path-constrained partial music synchronization. In *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, U.S.A., 2008, pp. 65–68. IEEE.

[52] M. Müller, B. Demuth, and B. Rosenhahn. An evolutionary approach for learning motion class patterns. In G. Rigoll, ed., *Pattern Recognition*, Munich, Germany, 2008, *LNCS 5096*, pp. 365–374. Springer.

[53] M. Müller and S. Ewert. Joint structure analysis with applications to music annotation and synchronization. In J. P. Bello, E. Chew, and D. Turnbull, eds., *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, Pennsylvania, USA, 2008, pp. 389–394.

[54] M. Müller, S. Ewert, and S. Kreuzer. Making chroma features more robust to timbre changes. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009. IEEE.

[55] T. Ritschel, T. Grosch, J. Kautz, and H.-P. Seidel. Interactive global illumination based on coherent surface shadow maps. In C. Shaw and L. Bartram, eds., *Proceedings Graphics Interface, May 2008*, Windsor, Ontario, Canada, 2008, ACM International Conference Proceeding Series, pp. 185–192. ACM.

[56] T. Ritschel, T. Grosch, M. H. Kim, H.-P. Seidel, C. Dachsbacher, and J. Kautz. Imperfect shadow maps for efficient computation of indirect illumination. In *Proceedings of ACM SIGGRAPH Asia 2008*, Singapore, 2008. ACM.

[57] B. Rosenhahn, T. Brox, D. Cremers, and H.-P. Seidel. Online smoothing for markerless motion capture. In F. Hamprecht, C. Schnörr, and B. Jaehne, eds., *Pattern Recognition*, Heidelberg, Germany, 2007, *LNCS 4713*, pp. 163–172. Springer.

[58] B. Rosenhahn, T. Brox, D. Cremers, and H.-P. Seidel. Modeling and tracking line-contrained mechanical systems. In G. Sommer and R. Klette, eds., *Robot Vision, 2nd International Workshop*, Auckland, New Zealand, 2008, *LNCS 4931*, pp. 98–110. Springer.

[59] B. Rosenhahn, T. Brox, and H.-P. Seidel. Scaled motion dynamics for markerless motion capture. In *2007 IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07. - Vol. 3*, Minneapolis, Minnesota, 2007, pp. 1203–1210. IEEE.

[60] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, D. Cremers, and H.-P. Seidel. Markerless motion capture of man-machine interaction. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

[61] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, and H.-P. Seidel. Staying well grounded in markerless motion capture. In G. Rigoll, ed., *Proceedings of the 30th DAGM Symposium*, München, Germany, 2008, *LNCS 5096*, pp. 385–395. Springer.

[62] W. Saleem, W. Song, A. Belyaev, and H.-P. Seidel. On computing best fly. In M. Sbert, ed., *Proceedings of the 23rd Spring Conference on Computer Graphics, 2007*, Budmerice, Slovakia, 2007, pp. 143–149. Comenius University.

[63] W. Saleem, D. Wang, A. Belyaev, and H.-P. Seidel. Automatic 2d shape orientation by example. In *2007 International Conference on Shape Modeling and Applications (SMI 2007)*, Lyon, France, 2007, pp. 221–225. IEEE.

[64] O. Schall, R. Zayer, and H.-P. Seidel. Controlled field generation for quad-remeshing. In E. Haines and M. McGuire, eds., *ACM Symposium on Solid and Physical Modeling 2008*, Stony Brook, New York, USA, 2008, pp. 295–300. ACM.

[65] C. Schmaltz, B. Rosenhahn, T. Brox, D. Cremers, J. Weickert, L. Wietzke, and G. Sommer. Region-based pose tracking. In J. Marti, J.-M. Benedi, A. Mendonca, and J. Serrat, eds., *Proceedings of the 3rd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2007)*, Girona, Spain, 2007, *LNCS 4478*, pp. 56–63. Springer.

[66] C. Schmaltz, B. Rosenhahn, T. Brox, J. Weickert, D. Cremers, L. Wietzke, and G. Sommer. Occlusion modeling by tracking multiple objects. In F. Hamprecht, C. Schnörr, and B. Jaehne, eds., *Pattern Recognition 2007*, Heidelberg, Germany, 2007, *LNCS 4713*, pp. 173–183. Springer.

[67] C. Schmaltz, B. Rosenhahn, T. Brox, J. Weickert, L. Wietzke, and G. Sommer. Dealing with self-occlusion in region based motion capture by means of internal regions. In F. J. López and R. B. Fisher, eds., *Proceedings of the 5th International Conference on Articulated Motion and Deformable Objects (AMDO 2008)*, Port d'Andratx, Mallorca, Spain, 2008, *LNCS 5098*, pp. 102–111. Springer.

[68] T. Schultz and H.-P. Seidel. Using eigenvalue derivatives for edge detection in DT-MRI data. In G. Rigoll, ed., *Pattern Recognition*, Munich, Germany, 2008, *LNCS 5096*, pp. 193–202. Springer.

[69] K. Shi, H. Theisel, H. Hauser, T. Weinkauf, K. Matkovic, H.-C. Hege, and H.-P. Seidel. Path line attributes - an information visualization approach to analyzing the dynamic behavior of 3D time-dependent flow fields. In H.-C. Hege, K. Polthier, and G. Scheuermann, eds., *Topology-Based Methods in Visualization II*, Grimma, Germany, 2009, Springer series of Mathematics and Visualization, pp. 75–88. Springer.

[70] W. Song, A. Belyaev, and H.-P. Seidel. Automatic generation of bas-reliefs from 3d shapes. In *2007 International Conference on Shape Modeling and Applications (SMI 2007)*, Lyon, France, 2007, pp. 211–214. IEEE.

[71] M. Sunkel, B. Rosenhahn, and H.-P. Seidel. Silhouette based generic model adaptation for marker-less motion capturing. In A. Elgammal, B. Rosenhahn, and R. Klette, eds., *Human Motion - Understanding, Modeling, Capture and Animation, Second Workshop, Human Motion 2007*, Rio de Janeiro, Brazil, 2007, *LNCS 4814*, pp. 119–135. Springer.

[72] A. Tevs, M. Bokeloh, M. Wand, A. Schilling, and H.-P. Seidel. Isometric registration of ambiguous and partial data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami Beach, Florida, USA, 2009. IEEE Computer Society.

[73] A. Tevs, I. Ihrke, and H.-P. Seidel. Maximum mipmaps for fast, accurate, and scalable dynamic height field rendering. In *Symposium on Interactive 3D Graphics (I3D 2008)*, Redwood City, California, USA, 2008, pp. 183–190. ACM.

[74] C. Theobalt, C. Rössl, E. de Aguiar, and H.-P. Seidel. Animation collage. In M. Gleicher and D. Thalmann, eds., *Symposium on Computer Animation*, San Diego, California, 2007, pp. 271–280. Eurographics.

[75] T. Thormählen, N. Hasler, M. Wand, and H.-P. Seidel. Merging of feature tracks for camera motion estimation from video. In *5th European Conference on Visual Media Production (CVMP 2008)*, London, UK, 2008. The Institution of Engineering and Technology.

[76] I. Wald, W. R. Mark, J. Günther, S. Boulos, T. Ize, W. Hunt, S. G. Parker, and P. Shirley. State of the art in ray tracing animated scenes. In D. Schmalstieg and J. Bittner, eds., *STAR Proceedings of Eurographics 2007*, Prague, Czech Republic, 2007, pp. 89–116. Eurographics Association.

[77] A. Yoshida, M. Ihrke, R. Mantiuk, and H.-P. Seidel. Brightness of the glare illusion. In S. Creem-Regehr and K. Myszkowski, eds., *Proceedings of ACM Symposium on Applied Perception in Graphics and Visualization*, Los Angeles, CA, USA, 2008, pp. 83–90. ACM.

[78] A. Yoshida, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Perceptual uniformity of contrast scaling in complex images. In C. Wallraven and V. Sundstedt, eds., *Proceedings of ACM Symposium on Applied Perception in Graphics and Visualization*, New York, USA, 2007, pp. 137–137. ACM.

[79] A. Yoshida, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Perception-based contrast enhancement model for complex images in high dynamic range. In B. E. Rogowitz and T. N. Pappas, eds., *Human Vision and Electronic Imaging XIII, IS&T/SPIE's 20th Annual Symposium on Electronic Imaging (2008)*, San Jose, CA, USA, 2008, vol. 6806, pp. 68060C–1–11. SPIE.

[80] S. Yoshizawa, A. Belyaev, H. Yokota, and H.-P. Seidel. Fast and faithful geometric algorithm for detecting crest lines on meshes. In M. Alexa, S. Gortler, and T. Ju, eds., *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, Maui, Hawaii, US, 2007, pp. 231–237. IEEE Computer Society.

[81] R. Zayer, B. Lévy, and H.-P. Seidel. Linear angle based parameterization. In A. Belyaev and M. Garland, eds., *Symposium on Geometry Processing*, Barcelona, Spain, 2007, pp. 135–141. Eurographics/ACM.

31 The Databases and Information Systems Group (D5)

31.1 Personnel

Director

Prof. Dr.-Ing. Gerhard Weikum

Researchers

Srikanta Bedathur Jagannath, PhD
Dr. Mouna Kacimi
Dr. Edwin Lewis-Kelham (January 2008–)
Dr. Arturo Mazeikas (March 2008–)
Dr. Thomas Neumann
Dr. Nicoleta Preda (September 2008–)
Maya Ramanath, PhD
Dr.-Ing. Ralf Schenkel (September 2002–November 2007, now Saarland University)
Dr. Mauro Sozio
Dr. Marc Spaniol (March 2008–)
Dr.-Ing. Fabian Suchanek (January 2009–)
Dr.-Ing. Martin Theobald (October 2008–)
Christos Tryfonopoulos, PhD (September 2006–December 2008, now University of Peloponnese)

PhD Students

Avishek Anand (January 2009–)
Ralitsa Angelova
Matthias Bender (–September 2007)
Klaus Berberich
Andreas Broschart (–November 2007, now Saarland University)
Tom Crecelius (–November 2007, now Saarland University)
Gerard de Melo
Dimitar Denev (July 2008–)
Laura Dietz (January 2009–)
Shady Elbassuoni (July 2007–)
Georgiana Ifrim (–January 2009, now University of Aarhus)
Gjergji Kasneci
Julia Luxenburger (–December 2008, now Google Zürich)

Sebastian Michel (–July 2007, now EPFL Lausanne)
Ndapandula Nakashole (March 2009–)
Hanglin Pan (–June 2008)
Lizhen Qu (February 2009–)
Fabian Suchanek (–December 2008)
Bilyana Taneva (July 2008–)
Yafang Wang (March 2009–)
Josiane Xavier Parreira
Christian Zimmer (–December 2008)

Associated Independent Research Group

Dr.-Ing. Ralf Schenkel (December 2007–)
Andreas Broschart (December 2007–)
Tom Crecelius (December 2007–)

Secretaries

Olha Condor (March 2008–March 2009)
Petra Schaaf

31.2 Visitors

In the time period from April 2007 to April 2009, the following researchers visited our group:

Bernhard Seeger	26.03.07–04.05.07	University of Marburg
Pratyus Patnaik	01.01.07–31.05.07	IIIT Allahabad
Sarath Kumar Kondreddi	03.09.07–	IIIT Allahabad
Paraskevi Raftopoulou	28.09.07–21.12.07 08.02.08–31.07.08	Technical University of Crete (TUC)
Qi Zhang	17.12.07–31.10.08	University of Science and Technology of China
Pierre Senellart	04.02.08–01.08.08	INRIA Futurs, Orsay
Divesh Srivastava	14.03.08–12.09.08	AT&T Labs-Research
Sihem Amer-Yahia	14.03.08	Yahoo! Research New York
Michal Shmueli-Scheuer	07.04.08–11.04.08	IBM Haifa
Arpit Singhal	01.05.08–25.07.08	CSE IIT Bombay
Alekhya Telekicherla	09.05.08–23.07.08	IIT Guwahati
Mounica Prodduturu	09.05.08–23.07.08	IIT Guwahati
Aparna Varde	14.05.08–14.07.08	Virginia State University
Xing Xie	27.06.08	Microsoft Research Asia
Konstantinos Morfonios	09.07.08–11.07.08	National and Kapodistrian University of Athens
Nicoleta Preda	10.07.08–11.07.08	University of Paris XI and INRIA
Nikos Mamoulis	01.09.08–31.05.09	University of Hong Kong

Ashwin Kumar Kayyoor	01.10.08–28.02.09	IIIT Allahabad
Manolis Koubarakis	29.10.08–31.10.08	National and Kapodistrian University of Athens
Eda Baykan	17.12.08–19.12.08	Ecole Polytechnique fédérale de Lausanne
Antti Ukkonen	18.01.09–20.01.09	Helsinki University of Technology
Gautam Parai	19.01.09–31.05.09	IIIT Allahabad
Leong Hou U	1.02.09–31.05.09	University of Hong Kong
Gaurav Pandey	01.02.09–31.07.09	ESCP-EAP, London
David Yang	18.02.09	University of Hong Kong
Alessandro Fiori	01.03.09–31.08.09	Politecnico di Torino
Djoerd Hiemstra	17.03.09	University of Twente
Royi Roonen	22.03.09	Technion, Israel

31.3 Group Organization

D5 has been established in October 2003. It is headed by Gerhard Weikum, and currently consists of 15 doctoral students and 12 post-doctoral researchers. The group's research aims to bridge the models, algorithmic methods, and architectural paradigms of the three fields of database systems, information retrieval, and data mining. The work is organized into five technical areas:

- knowledge harvesting (coordinated by Gerhard Weikum),
- Web and text mining (coordinated by Srikanta Bedathur),
- ranking and uncertain data management (coordinated by Martin Theobald),
- query processing and optimization (coordinated by Thomas Neumann), and
- distributed data and communities (coordinated by Mauro Sozio).

While each of these areas has its specific focus, they do have thematic overlaps and mutual synergies. Likewise, the students and researchers working in each area are dynamically formed teams rather than specifically dedicated staff. The area coordinators together serve as an internal steering board and individually as communication links between areas. They also play the role of co-advisors or de-facto advisors of various doctoral students.

In addition to the above five areas, D5 intensively cooperates with one of the independent research groups in the Excellence Cluster with focus on:

- efficient search in semistructured data spaces (headed by Ralf Schenkel, see Section 31.9).

31.4 Knowledge Harvesting

Coordinator: Gerhard Weikum

Universal, comprehensive knowledge bases have been an elusive AI goal for many years [5, 6]. Ontologies and thesauri such as OpenCyc, SUMO, WordNet, or UMLS (for the biomedical domain) are achievements along this route [8]. But they are typically focused on intensional knowledge about semantic classes. For example, they would know that mathematicians are scientists, that scientists are humans (and mammals and vertebrates, etc.); and they may also know that humans are either male or female, cannot fly, but can compose and play music, and so on. However, the currently available ontologies typically disregard the extensional knowledge about individual entities: instances of the semantic classes that are captured and interconnected in the ontology. For example, none of the above mentioned ontologies knows more than a handful of concrete mathematicians (or famous biologists etc.). Today, the best source for extensional knowledge is probably Wikipedia, providing a wealth of knowledge about individual entities and their relationships. But most of this knowledge is only latent, by being embedded in the natural-language text of Wikipedia articles or, in the best case, reflected in the semistructured components of Wikipedia like infoboxes and the category system.

The recent success of knowledge-sharing communities and the advances in automated information extraction (IE) [2, 7] from textual and semistructured Web sources (e.g., Wikipedia) have enabled large-scale harvesting of entity-relationship-oriented facts to build a new generation of much richer knowledge bases. IE technology comprises methods from pattern matching (e.g., regular expressions), linguistic analysis (e.g., part-of-speech tagging or dependency parsing), and statistical learning. Projects like DBpedia [1] (`http://dbpedia.org/`), YAGO [4, 9, 10], Freebase (`http://freebase.org/`), KnowItAll [3], or Intelligence-In-Wikipedia [11] have successfully created semantic databases with many millions of facts about entities (e.g., persons, companies, locations) and relationships (e.g., bornOnDate, marriedTo, isCEOof, headquarteredIn). Our own work on the YAGO knowledge base and the associated projects NAGA (for search and ranking) and SOFIE (for advanced extraction and cleaning) are at the forefront of this exciting research avenue.

Our long-term agenda is twofold: develop general methodologies for knowledge harvesting, and demonstrate their practical viability and usefulness by constructing and maintaining specific knowledge bases from a variety of sources. This line of research includes the following achievements and ongoing projects. YAGO is a large collection of entities and relational facts that are harvested from Wikipedia and WordNet with high accuracy and reconciled into a consistent RDF-style "semantic graph" (see Subsection 31.4.1). The knowledge base (and also the extraction software) is publicly available and has been used in other knowledge-sharing projects including DBpedia. For further growing YAGO from Web sources while retaining its high quality, pattern-based extraction is combined with logic-based consistency checking in a unified framework coined SOFIE (see Subsection 31.4.2). For tapping on information that cannot be crawled and is accessible only via Web services, we are investigating the integration of dynamically invoked service functions with explicitly harvested prior knowledge (see Subsection 31.4.3). To connect entities, facts, and their terminologies in different languages, we are developing methods towards multilingual knowledge harvesting (see Subsection 31.4.4).

Finally, to capture the dynamics of evolving knowledge, we have started investigating the extraction and management of time-varying relations (e.g., former and current spouses and the corresponding time periods; see Subsection 31.4.5).

Once a rich knowledge base has been constructed, further issues arise about querying the knowledge and ranking search results for user-friendly knowledge discovery. This in turn poses challenges for efficient and scalable query processing on schema-less entity-relationship graphs. The methods for addressing these issues fall into the areas "Ranking and Uncertain Data" and "Query Processing and Optimization", and are discussed there, particularly, in Subsections 31.6.2 and 31.7.1.

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. F. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, eds., *ISWC/ASWC*, 2007, *LNCS 4825*, pp. 722–735. Springer.

[2] A. Doan, L. Gravano, R. Ramakrishnan, and S. Vaithyanathan, eds. *Special Issue on Managing Information Extraction.* ACM SIGMOD Record 37(4), 2008.

[3] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the Web. *Commun. ACM*, 51(12):68–74, 2008.

• [4] G. Kasneci, F. Suchanek, M. Ramanath, and G. Weikum. The YAGO-NAGA approach to knowledge discovery. *SIGMOD Record Special Issue on Managing Information Extraction*, 37(4):41–47, 2008.

[5] D. B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):32–38, 1995.

[6] D. B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project.* Addison-Wesley Pub (Sd), 1990.

[7] S. Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3):261–377, 2008.

[8] S. Staab and R. Studer, eds. *Handbook on Ontologies.* International Handbooks on Information Systems. Springer, 2004.

• [9] F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 697–706. ACM.

• [10] G. Weikum, G. Kasneci, F. Suchanek, and M. Ramanath. Database and information-retrieval methods for knowledge discovery. *Communications of the ACM*, 52(1), 2009.

[11] D. S. Weld, F. Wu, E. Adar, S. Amershi, J. Fogarty, R. Hoffmann, K. Patel, and M. Skinner. Intelligence in Wikipedia. In D. Fox and C. P. Gomes, eds., *AAAI*, 2008, pp. 1609–1614. AAAI Press.

31.4.1 YAGO: Fact Extraction from Wikipedia

Investigators: Fabian Suchanek, Gjergji Kasneci, and Gerhard Weikum

Ontological background knowledge can be a great asset to information technology. This applies, above all, to applications in the vision of the Semantic Web, but also to many other fields such as machine translation, query expansion, document classification, record linkage, and information integration in general. Previous approaches for ontology construction have often traded coverage against accuracy. Manually created ontologies have high accuracy, but are expensive to build and thus have low coverage. Automatically constructed ontologies, on the other hand, can achieve high coverage, but often exhibit poor quality by including a non-negligible fraction of spurious facts.

To address these issues, we have developed the ontology YAGO[1] [5, 6]. YAGO combines high coverage with high quality. Its core is assembled from one of the most comprehensive lexicons available today, Wikipedia. But rather than using natural language processing on the articles of Wikipedia, our approach builds on Wikipedia's *infoboxes* and *category pages*. Infoboxes are standardized tables that contain basic information about the entity described in the article. For example, there are infoboxes for countries, which contain the native name of the country, its capital, and its size. Infoboxes are much easier to parse and exploit than natural language text. Category pages are lists of articles that belong to a specific category (e.g., Elvis is in the category of American rock singers). These lists give us candidates for entities (e.g., *Elvis*), candidates for concepts (e.g., *IsA(Elvis, rockSinger)*), and candidates for relations between entities (e.g. *nationality(Elvis, American)*). The instances of relationships between entities are also referred to as facts.

In an ontology, concepts have to be arranged in a taxonomy (class hierarchy) for proper usage. The Wikipedia categories are indeed arranged in a hierarchy, but this hierarchy is barely useful for ontological purposes. For example, Elvis is in the super-category named "Grammy Awards", but Elvis is a Grammy Award *winner* and not a Grammy Award. WordNet [2], in contrast, provides a clean and carefully assembled hierarchy of thousands of concepts. But the Wikipedia concepts have no obvious counterparts in WordNet. We developed techniques that link the two sources with near-perfect accuracy. Our method is the first approach that accomplishes this unification between WordNet and facts derived from Wikipedia with an accuracy of 97%. This allows the YAGO ontology to profit, on one hand, from the vast amount of individuals known to Wikipedia, while exploiting, on the other hand, the clean taxonomy of concepts from WordNet. Currently (as of March 2009), YAGO contains 2 million entities and 20 million facts about them. Figure 31.1 shows a small excerpt of YAGO in pictorial form, to give a visual impression of its structure and wealth.

We enforce the high accuracy of our extraction heuristics through *type checking* [4]. Type checking leverages the information that has already been extracted to verify the plausibility of newly extracted data. Type checking can be used both in a *reductive fashion*, by eliminating facts that are implausible, and in an *inductive fashion*, by adding supplemental facts so that the ontology becomes consistent. We have conducted an extensive evaluation study, which shows that YAGO has an overall accuracy of at least 95%.

YAGO is based on a data model of entities and binary relations. But by means of reification

[1] Yet Another Great Ontology

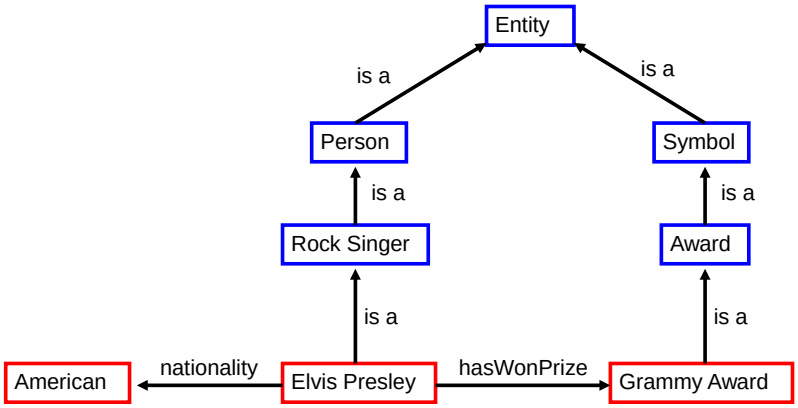


Figure 31.1: Excerpt of the YAGO Knowledge Base

(i.e., introducing identifiers for relation instances) we can also express relations between facts (e.g., which facts were found on which Web site), n-ary relations (e.g. that *Elvis* won the *Grammy Award* in *1967*), and general properties of relations (e.g., transitivity or acyclicity). This entails that, despite its expressiveness, the YAGO data model is still decidable and can be mapped to the W3C data model RDFS. Furthermore, we have developed a query language as a natural extension of our data model, which allows querying reified facts. YAGO is freely available at `http://mpi-inf.mpg.de/yago/`; this Web site also features an online query interface. The YAGO ontology has been made available for and integrated into other knowledge-sharing projects, including DBpedia [1] (`http://dbpedia.org/`), UMBEL (`http://umbel.org/`), Freebase (`http://freebase.com/`), and SUMO [3] (`http://www.ontologyportal.org/`).

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. F. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, eds., *ISWC/ASWC*, 2007, *LNCS 4825*, pp. 722–735. Springer.

[2] C. Fellbaum, ed. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

- [3] G. de Melo, F. Suchanek, and A. Pease. Integrating YAGO into the Suggested Upper Merged Ontology. In *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008)*, Dayton, OH, USA, 2008, vol. 1, pp. 190–193. IEEE Computer Society.
- [4] F. Suchanek. *Automated Construction and Growth of a Large Ontology*. Phd thesis, Universität des Saarlandes, 2008.
- [5] F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 697–706. ACM.
- [6] F. Suchanek, G. Kasneci, and G. Weikum. YAGO - a large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 6(3):203–217, 2008.

31.4.2 SOFIE: Fact Extraction from Web Sources

Investigators: Fabian Suchanek, Mauro Sozio, and Gerhard Weikum

Recently, several projects such as YAGO [5] (see also Section 31.4.1), DBpedia [1], KnowItAll [3], and Intelligence-In-Wikipedia [7] have automatically constructed large knowledge bases by using information extraction (IE) on encyclopedic sources and Web pages. The projects that were able to build large collections of ontology-style quality have particularly focused on extracting information from the semi-structured components of Wikipedia (such as infoboxes and the category system). In order to achieve an even broader coverage, additional Web sources must be tapped on. The potentially richest asset are natural-language documents, such as news articles, biographies, scientific publications, and also the full text of Wikipedia articles. But so far even the best IE methods have typically achieved only 80 percent accuracy (or less) in such settings (depending on the difficulty of the target relations to be extracted). While this may be good enough for some applications [2], the error rate is unacceptable for an ontological knowledge base that should form the basis of semantic reasoning.

Extracting new facts that are consistent with an existing ontology entails several, highly intertwined problems:

- **Pattern selection:** Facts are extracted from natural language documents by finding meaningful patterns in the text. The accuracy of this technique critically depends on having a variety of meaningful patterns. It can be further boosted if counter-productive patterns are systematically excluded. Thus, discovering and assessing patterns is a key task of IE.
- **Entity disambiguation:** For ontological IE, the words or phrases from the text have to be mapped to entities in the ontology. In many cases, this mapping is ambiguous. The word "Paris", for example, can denote either the French capital or a city in Texas or the character from the Iliad or the first name of the celebrity Paris Hilton. Since many location names, companies, or product names are ambiguous, finding the intended meaning of a word is often a difficult task.
- **Consistency checking:** The newly extracted facts have to be logically consistent with the existing ontology. Consistency checking is a difficult problem by itself. In our case, the problem is particularly challenging, because a large set of IE-provided noisy candidates has to be scrutinized against a trusted core of facts.

We have developed a new approach to these problems. Rather than addressing each of them separately, we provide a unified model for ontology-oriented IE that solves all three issues simultaneously [4, 6]. To this end, we cast known facts, hypotheses for new facts, word-to-entity mappings, gathered sets of patterns, and semantic consistency constraints into a unifying framework of logical clauses. This way, all three problems boil down to checking one large set of clauses. Solving the problems means finding the literals (i.e., primitive clauses that contain only one predicate symbol with constants as arguments) that have to be set to `true` so that a maximum number of (weighted) clauses is satisfied, or the total weight of satisfied clauses is maximized. This problem is known as the *Weighted Maximum Satisfiability problem* (Weighted MAX-SAT).

The Weighted MAX-SAT problem is NP-hard. Our instance is particularly large, as we use the large YAGO ontology [5] as prior knowledge and every fact in YAGO forms a literal. This is why we have to resort to an approximation algorithm. Our problem contains many *functional dependencies*, i.e., clauses that enforce the right-hand uniqueness of relations. This entails that one of the best approximation algorithm for Weighted MAX-SAT, Johnson's Algorithm, has a particularly bad result quality. Therefore, we have developed a new algorithm that is tailored to our setting, the *Functional MAX-SAT Algorithm* (FMS Algorithm). We have proven that the FMS Algorithm has a worst-case approximation guarantee of 1/2. We show by experiments that the algorithm performs much better in practice and even computes the optimum in many cases.

We have implemented the algorithm in a system called SOFIE[2]. SOFIE parses the input documents and produces literals for the patterns and fact hypotheses that occur in the text, along with occurrence statistics. These are then substituted into the constraints on the relations of interest and form a large set of logical clauses with weights derived from the occurrence statistics. Next, SOFIE runs the FMS Algorithm on these clauses and on the literals from the existing YAGO ontology. Last, we assign truth values to the literals that appear in the clauses, and those literals that are assigned the value `true` are output as new facts. We have conducted experiments on a variety of real-life textual and semi-structured sources (e.g., biographies from the Web). SOFIE requires about one day to process several thousand Web documents (on a standard PC), and produces high-quality output. It consistently achieves a precision of over 90%. This shows the practical viability of our unifying SOFIE approach.

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. F. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, eds., *ISWC/ASWC*, 2007, *LNCS 4825*, pp. 722–735. Springer.

[2] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the Web. *Commun. ACM*, 51(12):68–74, 2008.

[3] O. Etzioni, M. J. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Unsupervised named-entity extraction from the Web: An experimental study. *Artif. Intell.*, 165(1):91–134, 2005.

- [4] F. Suchanek. *Automated Construction and Growth of a Large Ontology.* Phd thesis, Universität des Saarlandes, 2008.

- [5] F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 697–706. ACM.

- [6] F. Suchanek, M. Sozio, and G. Weikum. SOFIE: A self-organizing framework for information extraction. In *Proceedings of the 18th World Wide Web Conference (WWW 2009)*, Madrid, Spain, 2009. ACM.

[2] Self-Organizing Framework for Information Extraction

[7] F. Wu and D. S. Weld. Automatically refining the Wikipedia infobox ontology. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, eds., *WWW*, 2008, pp. 635–644. ACM.

31.4.3 ANGIE: Active Knowledge Services

Investigators: Nicoleta Preda, Fabian Suchanek, and Gjergji Kasneci

Knowledge bases such as YAGO [3] (see Subsection 31.4.1) contain rich information about millions of entities and relationships between them, represented, for example, in the Semantic-Web data model RDF. Systems of this kind provide powerful search capabilities by SPARQL-like query languages and user-friendly ways of knowledge exploration. However, such a knowledge base can never be complete and inevitably exhibits gaps that may irritate the user during interactive access. For example, suppose a user finds the biography of a singer interesting and then wants to find all songs and albums by this singer, including the latest ones. Crawling additional Web sites on music and extracting the missing data is often infeasible because of site restrictions and because the site's information is continuously changing. Moreover, some knowledge needs are inherently ephemeral: for example, asking for the current rating of a movie (by averaging user reviews) or the chart rank of a song. Fortunately, there is an increasing number of Web services on music, movies, books, business directories, etc. that could potentially fill the gaps in the database.

The ANGIE project [2], aims at building a system that can retrieve data from Web services on the fly, whenever the local knowledge base does not suffice to answer a user query. In ANGIE, Web services act as dynamic components of the knowledge base that deliver knowledge on demand. To the user, this is fully transparent; the dynamically acquired knowledge is presented as if it were stored in the local knowledge base. We have developed a model for declarative definition of functions embedded in the local knowledge base. The (REST or WSDL) interfaces of available Web services are cast into such functions. Parameter bindings are automatically constructed by ANGIE, services are invoked, and the semistructured information returned by the services are dynamically integrated into the knowledge base.

Our approach is to treat the functions as intensional components of the knowledge base and represent them within the knowledge base itself. Executing a function provides bindings for its output parameters, filled in on demand by the invoked Web services. These output are added to the knowledge base and can be directly read later, for evaluating other user queries. A function definition is a parameterized query, represented in the form of a semantic graph (edges are relationship names and the nodes are variables or entities). Furthermore, the edges are partitioned into input preconditions and output postconditions. The preconditions have to be fulfilled before the function can be called (by appropriate parameters filled in by ANGIE before external services are invoked). The postconditions correspond to output parameters of the invoked services. This representation naturally extends the model of semantic knowledge bases.

We have developed a *query rewriting algorithm* that determines one or more function composition that need to be executed in order to evaluate a SPARQL-style user query. The key idea is that the local knowledge base can be used to guide the selection of values used as input parameters of function calls. This is in contrast to the conventional approaches in the literature which would exhaustively materialize all values that can be used as binding values for the input parameters. The algorithm first generates a set of function instantiations.

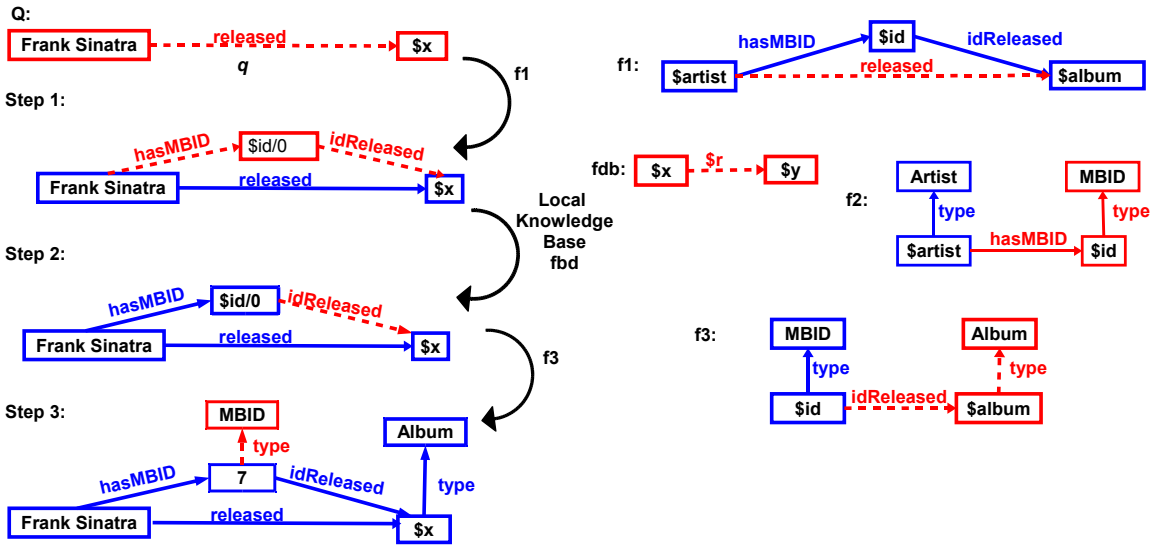


Figure 31.2: Query rewriting (left) and function definitions (right).

A function instantiation binds variables of the function definition with entities (possibly returned by previously executed function instantiations). The algorithm takes as input a set function definitions. First, it marks all the edges in the query as unsolved edges. The algorithm implements a recursive search on all possible rewritings into function compositions. In each recursion step, a function whose postcondition can cover unsolved edges in the query, is chosen. The function instantiations are sorted according to an estimation of their benefit/cost ratio (based on quality-of-service properties of the underlying services). The most promising option is selected, and the chosen function is appended to the query. The newly solved edges are marked as solved; the preconditions of the newly appended function are marked as unsolved. When all the edges become solved, the algorithm has found a solution and outputs the function composition that must be executed. We have proven the completeness of the rewriting algorithm: if there is a sequence of function calls that produces an answer for the user query, the algorithm computes it.

Consider, for instance, the query Q in Figure 31.2, and the function definitions defined in the left side of the figure. Unsolved edges are depicted as dotted lines. Q asks for the albums released by Frank Sinatra, function f_2 maps an artist to its id in the MusicBrainz world, f_3 maps the id to the albums released by the artist, and f_1 is a function-composition rule equivalent to the Horn clause:

$$released(\$artist, \$album) \leftarrow hasMBID(\$artist, \$id) \wedge isReleased(\$id, \$album)$$

The algorithm searches a substitution that satisfies q. In our case, f_1 can satisfy q with $\sigma(\$artist)$=*FrankSinatra* and $\sigma(\$album)$=$x. Hence, q is marked as solved, and f is appended to the query. The preconditions of f become new members of Q. In the next steps, the algorithm tries to satisfy the edges marked with *hasMBID* and *idReleased*. The algorithm chooses the local function f_{db} to satisfy the edge marked with *hasMBID*, as an id for *Frank Sinatra* is already stored in the local knowledge. Note that another choice is f_2. But this would

lead to a more costly solution, as f_2 is executed remotely. In the next step, the algorithm chooses f_3 to satisfy the edge marked with *idReleased*. The result is a rewriting that "covers" the initial query completely.

A prototype system has been built, and we are conducting an experimental evaluation. Our system consists of several components: (*i*) *a rewriting module* implementing the algorithm outlined above, (*ii*) *the existing RDF-3X engine* [1] that carries out the execution of the rewritings and uniformly integrates the data from the local knowledge base and the data returned by the Web service calls, (*iii*) *a mapping tool* that maps the XML schemas of the XML data returned as service results onto the RDF model of the knowledge base. A more detailed description of the ANGIE project is given in [2].

References

- [1] T. Neumann and G. Weikum. RDF-3X: a RISC-style engine for RDF. *Proceedings of the VLDB Endowment*, 1(1):647–659, 2008.
- [2] N. Preda, F. Suchanek, G. Kasneci, T. Neumann, and G. Weikum. Coupling knowledge bases and Web services for active knowledge. Research Report MPI-I-2009-5-004, Max Planck Institute for Informatics, Saarbrücken, Germany, 2009.
- [3] F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 697–706. ACM.

31.4.4 Multilingual Knowledge Harvesting

Investigators: Gerard de Melo and Gerhard Weikum

With the enormous growth of internet penetration all over the world, the English language represents a constantly decreasing fraction of the Web. China and the EU each have greatly surpassed the U.S. in the number of Internet users, and other regions are expected to follow. Multilingual knowledge bases address this development by providing entity labels in multiple languages and making the semantic connections between words and names in different languages explicit. Such information can be useful for various forms of natural language processing, information retrieval, artificial intelligence, as well as human consultation.

To create such resources efficiently, we have developed techniques to automatically transfer monolingual knowledge to other languages. There are several assets to be considered for this purpose. For individual entities such as cities and products, multilingual labels can be derived from the "interwiki" links present in Wikipedia. For example, "Roma", "Rome", and "Rom" denote the Italian capital in three different languages (Italian, English, and German), but "Roma" is used in German with a very different meaning, referring to an Eastern European ethnic group. For upper-level concepts like semantic classes (e.g., "soccer player"), as well as for verbal and adjectival concepts, similarly rich sources do not exist. The best starting point is probably WordNet [1], a large thesaurus that maps *words* onto *word senses* (concepts), or *senses* for short. This mapping considers both synonymy (different words with the same sense) and homonymy (same word used for different senses), and WordNet also considers hyponymy/hypernymy (subclass/superclass) and meronymy/holonymy (part-of/composition)

relations among senses. However, the original WordNet covers only the English language. There are a few dozens of non-English versions of WordNet, but they are much smaller than the English version and thus are far from being similarly useful, and, of course, hundreds of languages are totally disregarded.

One can use word-level translation dictionaries to create links from the English WordNet to a previously non-existing or very sparse non-English WordNet. However, a straightforward approach runs into major difficulties because of synonyms and homonyms. For example, a word such as "bat" has 10 senses in the English WordNet, but a German translation such as "Fledermaus" (the animal) only applies to a small subset of those senses. Conversely, the English word "faucet" (or "tap") would be translated into the German word "Hahn", which is highly ambiguous in German as the same word is used for "rooster".

We approached this challenge by harnessing machine-learning techniques [3, 4]. An initial knowledge graph G_0 is constructed by extracting information from the English WordNet, translation dictionaries, thesauri (such as the GEMET thesaurus), ontologies, and parallel corpora (e.g., the OpenSubtitles corpus), and then applying heuristics to increase the density of the graph and merge edges. The knowledge graph contains nodes for words and phrases (e.g., composite nouns that denote names such as "United States of America") as well as sense nodes representing word senses. A sequence of graphs G_i is iteratively derived by predicting new edge weights for semantic links between words and sense nodes using RBF-kernel SVM models. The models are learned from a training set of labeled links between words and senses using a number of graph-based statistical scores that represent properties of the previous graph G_{i-1} and additionally make use of measures of semantic relatedness and corpus frequencies. Relying on multiple iterations allows us to draw on multilingual evidence for greater precision and recall, due to mutual reinforcement and propagation effects. For instance, in the first iteration, we could determine that the German word "Fledermaus" is linked to the animal sense of "bat" with high probability, and then in the next iteration this may help us infer that the Turkish word "yarasa" has the same meaning.

We have successfully applied these techniques to create a large-scale multilingual thesaurus, which, at a reasonably high level of precision (85-90%), provides around 1,000,000 word-sense links for over 600,000 words in a multitude of languages. The coverage goes far beyond that of existing multilingual resources such as hand-crafted thesauri. These figures can easily grow even further as we enlarge the input graph (by tapping on additional sources). The structure, too, is reasonably rich, including hyponymy/hypernymy and several other relations.

Additional experiments have shown that such automatically built WordNets are beneficial in a variety of application tasks. Examples we studied include cross-lingual text classification and semantic-relatedness estimation, where we outperform high-quality manually created resources [2].

References

[1] C. Fellbaum, ed. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

- [2] G. de Melo and G. Weikum. On the utility of automatically generated wordnets. In A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, and P. Vossen, eds., *Fourth Global WordNet Conference (GWC 2008)*, Szeged, Hungary, 2007, pp. 147–161. University of Szeged.

- [3] G. de Melo and G. Weikum. A machine learning approach to building aligned wordnets. In *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, Hong Kong, 2008, pp. 163–170.
- [4] G. de Melo and G. Weikum. Towards a universal wordnet by learning from combined evidence. Research Report MPI-I-2009-5-005, Max Planck Institute for Informatics, Saarbrücken, Germany, 2009.

31.4.5 Temporal Fact Extraction

Investigators: Qi Zhang, Fabian Suchanek, and Gerhard Weikum

Information extraction technology is often used as if the underlying world were time-invariant. When information sources such as Web pages are processed a second time, the extracted data is simply added to an existing knowledge base or overwrites prior knowledge. This is appropriate for some forms of extraction tasks, e.g., finding birthdates of famous people, but inappropriate for evolving facts, e.g., presidents of countries or CEOs of companies. In fact, time-dependent relations seem to be far more common than time-invariant ones. For example, finding all spouses of famous people, current and former ones, involves understanding temporal relations.

In our overriding theme of knowledge harvesting, we have started looking into the issues of time-dependent facts and how to extract them and represent them in the knowledge base [2]. Using reification of binary relations, we have added a layer of time annotations to our knowledge model: each fact can be associated with a begin and an end timestamp. The timestamps themselves are actually intervals for denoting lower (earliest) and upper (latest) bounds so as to reflect uncertainty. This way we can handle different time granularities in a uniform manner; for example, knowing that Angela Merkel became the chancellor of Germany in 2005, but not knowing the exact date, can be represented by January 1, 2005 as a lower bound and December 31, 2005 as an upper bound. And we can easily represent the yet undefined end point by stating *now* and *infinity* as lower and upper bound, respectively. As we obtain more precise information in subsequent extractions, we can refine these intervals. Moreover, we can use logical reasoning to adjust, refine, or correct the temporal knowledge. For example, we can use birthdates or other key dates (e.g., graduation) as bounds for other events (e.g., joining a university as a professor).

For the extraction of temporal information associated with binary relations, we extended our former work on the LEILA tool which combines deep linguistic analysis (by a dependency-grammar parser) with seed-based statistical learning [1]. We have applied these techniques to sources like Wikipedia infoboxes, Reuter's news feeds, and news pages provided by Google, in order to extract time-annotated facts about companies acquiring other companies and company CEOs. Our experimental evaluation showed that we can obtain fairly high precision for the extracted information, but these are still very preliminary findings.

References

[1] F. Suchanek, G. Ifrim, and G. Weikum. Combining linguistic and statistical analysis to extract relations from Web documents. In T. Eliassi-Rad, L. Ungar, M. Craven, and D. Gunopulos, eds., *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining (KDD 2006), Philadelphia, PA, USA, 2006, pp. 712–717. ACM. Acceptance Ratio 1:5.

- [2] Q. Zhang, F. Suchanek, and G. Weikum. TOB: Timely ontologies for business relations. In *11th International Workshop on Web and Databases 2008 (WebDB 2008)*, Vancouver, Canada, 2008. ACM.

31.5 Web and Text Mining

Coordinator: Srikanta Bedathur

Increasing amounts of information pertaining almost all walks of life are published, archived and made accessible over the Web. In addition, the advent of Web 2.0 technologies has led to increased interactions, dynamics and diversity of content. Our goal is to detect and analyze patterns, trends, and salient properties in this huge wealth of information. This leads us to tackle a variety of research challenges, including, high-quality content classification, summarization of semi-structured content, learning from user preferences, predictions in complex social networks, understanding the evolution of information, and quantifying the coherence of information archives.

31.5.1 Indexing, Mining, and Querying the Evolving Web

Investigators: Srikanta Bedathur, Klaus Berberich, Thomas Neumann, and Gerhard Weikum in cooperation with Nikos Mamoulis and Jens Dittrich

For the Web, maybe the only thing constant about it is the fact that it always evolves. In addition to an almost obvious growth of the Web through the infusion of fresh content, there are many sources of Web evolution. First, thanks to improved digitization techniques, information that was originally published long before the Web's inception is made available in digital form. Prominent examples include the newspaper archives such as those of New York Times (`http://www.nytimes.com/`), Time Magazine (`http://www.time.com/`), Süddeutsche Zeitung (`http://www.sueddeutsche.de/`), etc. Second, pure born-digital content – such as the Wikis, Blogs, discussion forums, and the Web itself – are constantly modified, expanded, and kept up-to-date. Some collections such as Wikipedia (`http://www.wikipedia.org/`) maintain the explicit history of these changes, and others like the Internet Archive (`http://www.archive.org/`) have implicit time associated with their contents. Thus, the Web has turned into a collective long-term memory of mankind, rich not only by content, but also by its even richer evolutionary history that potentially reflects the evolution of human knowledge itself.

However, simple amassing of information is not an end in itself. We have to make sure that future generations can still access, interpret, and extract useful knowledge from these sources. Achieving this requires us to overcome challenges in issues such as efficient search, scalable mining and semantic interpretability of results.

Indexing the Evolving Content

Large versioned corpora offer a great potential for enhancing a variety of intelligence gathering activities that can utilize the evolutionary history of content. In order to unleash this potential,

it is imperative to have sophisticated search and mining capabilities over them. Text search on currently available large-scale evolving collections is still mostly *time-agnostic*. On the one end of the spectrum, we see approaches like those of Wikipedia search, which index only the most recent version of documents. On the other end lie systems such as NutchWAX (`http://archive-access.sourceforge.net/projects/nutch/`) which treat each version as a separate document, thus potentially diluting the query results with many near-identical document versions.

Although such approaches are sufficient for simple information needs, they are neither appropriate nor adequate for advanced needs. For example, consider the following queries:

- In preparation for a documentary, a journalist needs to research changing political and societal opinions about the war in Iraq which began in 2003.
- First reviews about the movie "Harry Potter and the Philosopher's Stone" that was released on 22 Nov. 2001.
- Gather Web pages about George W. Bush during the first year and last year of his presidency.

Such information needs cannot be effectively satisfied using standard search engines, since many of the relevant Web pages have disappeared and results are therefore dominated by more recent content. Although archives preserve the relevant Web pages, their poor accessibility make it extremely difficult, if not impossible, to locate them. Given that the targeted data sets are in the order of Terabytes (as the complete revision history of English Wikipedia) or even Petabytes (as the Internet Archive), naïve approaches to time-travel text search fail to scale and do not provide quick response times as today's users are accustomed to.

In order to serve such historical information needs effectively, we introduce the notion of *time-travel text search* over these versioned collections. Formally, a time-travel query $\mathcal{Q}$ is defined as a pair $\langle Q_{ir}, Q_{tc} \rangle$, where Q_{ir} is the IR-style keyword query and Q_{tc} is the target temporal context. This *time-travel query* is then evaluated over the state of the temporally versioned text collection as of time period defined by Q_{tc}.

In [2, 3, 4, 5], we developed a scalable framework called TTIX for indexing large-scale evolving text collections and efficiently answering time-travel queries. Our approach builds on the highly scalable inverted-file index that has become the de facto standard for indexing large-scale text. First, we extend the inverted file index to include the temporal variations in the tf and *idf*-values needed for relevance ranking. Since *idf*-values are time-varying with respect to the whole collection, we manage them as a separate time series data. The tf-values, however, are integrated into the payload of inverted list entries as follows:(d, p, $[t_b, t_e)$). Note that there is one posting for every version of each document in the collection.

Next, we control the resulting index-size blowup by means of *approximate temporal coalescing*. This tunable technique capitalizes on the high redundancy in successive versions of the same document. It reduces the number of postings in an index list by merging sequences of temporally consecutive postings of a document that have almost equal payloads, while keeping the maximal error made bounded by the tuning parameter ϵ. We adapt techniques proposed for piecewise constant representation of a time-series [8], and for generating a

compact relative-error histogram [7]. Experiments have shown that this step can reduce the index size by more than 60% with negligible effect on the top-10 query results.

While the previous step reduces the index size significantly, during query processing many postings have to be read even though they do not qualify for the given Q_{tc}, thus adversely affecting the query processing efficiency. Our third step tackles this problem, by introducing the notion of *sublist materialization*, based on the idea that query processing can be sped up if shorter index lists that contain information valid for a contiguous time interval are materialized separately. If we take this idea to its extreme, we obtain a *performance optimal* sublist materialization where a separate inverted list is maintained for each unique interval (also called as a *elementary interval*) defined by time-stamps in the list – but at the cost of enormous increase in the index size. This space blowup occurs due to the fact that coalesced postings that span different subintervals chosen for materialization are replicated in each of these materializations leading to a blowup in the index size. Thus there is a tension between the effective index size and the gain in performance due to materialization. We resolve this via two principled optimization problem formulations, which we call *Performance-Guarantee* and *Space-Bound* sublist materializations.

The first formulation, the performance-guarantee approach, finds a materialization $\mathcal{M}$ that requires minimal amount of space, while guaranteeing for any time-travel query that performance is worse than optimal by at most a factor of $\gamma \geq 1$. Formally,

$$\underset{\mathcal{M}}{argmin}\ S(\,\mathcal{M}\,) \quad \text{s.t.}$$

$$\forall\,[t_i,\,t_{i+1})\ \in\ \mathcal{E}\ :\ PC(\,[t_i,\,t_{i+1})\,)\ \leq\ \gamma \cdot |L_v : [t_i,\,t_{i+1})|\ ,$$

where $S(\,\mathcal{M}\,)$ is the size of the materialization $\mathcal{M}$, $\mathcal{E}$ is the set of all *elementary intervals* in the list, $PC(\,[t_i,\,t_{i+1})\,)$ is the processing cost of a query q^t for $t \in [t_i,\,t_{i+1})$, and $|L_v : m|$ represents the size of an individual sublist $m \in \mathcal{M}$. An optimal solution to this problem can be computed by means of *induction*, leading to algorithm with a time-complexity in $O(n^2)$. Taking the dual view, the space-bound approach constrains the index size blowup due to replication by $\kappa \geq 1$, and generates a materialization with optimal expected processing cost. This problem can be formally stated as,

$$\underset{\mathcal{M}}{argmin} \sum_{[t_i,\,t_{i+1})\,\in\,\mathcal{E}} P(\,[t_i,\,t_{i+1})\,) \cdot PC(\,[t_i,\,t_{i+1})\,) \quad \text{s.t.}$$

$$\sum_{m \in \mathcal{M}} |L_v : m| \leq \kappa\,|L_v|\ ,$$

where $P(\,[t_i,\,t_{i+1})\,)$ represents the probability of a query q^t with $t \in [t_i,\,t_{i+1})$. The resulting optimization problem can be optimally solved by dynamic programming over time axis of the inverted list. Unfortunately, the algorithm has time complexity in $O(n^3|L_v|)$ and space complexity in $O(n^2|L_v|)$, which makes it impractical for large datasets. In practice, an approximate solution can be obtained efficiently using *simulated annealing* technique in $O(n^2)$.

Experiments on large text collections, such as the entire version history of Wikipedia, have shown that these two techniques can achieve close-to-optimal query performance of time-travel queries while inducing only a small blowup of the index size.

Querying the Past

While TTIX provides an efficient indexing solution for time-travel queries, it does not enable semantic interpretability of queries and results. While searching document archives such as the newspaper archives containing articles that were published a long time ago, one has to deal with the growing gap between the terminology used in old documents and the terminology that users employ to formulate queries.

To illustrate this point, consider a query saint petersburg architecture. When using state-of-the-art retrieval techniques, old but possibly relevant documents about buildings in *Leningrad* are most likely to be missed. This is because the retrieval systems are unaware of the fact that both *Leningrad* and *Saint Petersburg* refer to the same city, but in different eras. As another example, consider the query ipod hearing damage for which articles relevant for walkman hearing damage from earlier times can also be considered equally relevant.

Our approach to solving this problem is to *reformulate* the queries issued by users, so that old but relevant documents that contain different terminology are retrieved. To this end, we first partition the archive into temporally coherent collections, e.g., groups of documents published in the same year. We compute term co-occurrence statistics for each of these partitions, which provides us with a model of the language use during corresponding period. As concrete examples, we observe frequent co-occurrence for the terms leningrad and hermitage in documents that were published in the 1980s – for more recent documents published in the 2000s, on the other hand, we observe that the terms saint petersburg and hermitage co-occur often together. Using these statistics, we can form time-specific term contexts as those terms that co-occur frequently with a specific term in documents that belong to a temporal partition.

In order to actually reformulate a given user query, the time-specific contexts are leveraged in the following way. Building on the idea that two terms are similar if they are used in context with similar terms, by comparing time-specific term contexts, we can estimate the semantic similarity of two terms when used at respectively different times. For our earlier example, from the fact that leningrad and saint petersburg co-occur frequently with terms such as moscow, peter the great, and hermitage in the 1980s and 2000s, we infer that leningrad when used in the 1980s is semantically similar to saint petersburg when used in the 2000s. When reformulating a given user query, we consider such terms that have a high degree of semantic similarity with the query terms as potential replacements. Query terms, though, are rarely independent and should thus not be replaced independently. Hence, we may have identified the terms smell and destruction as terms being semantically similar to hearing and damage, respectively. Clearly, putting together the terms walkman, smell, and damage as a reformulation of our initial query is nonsensical. To determine whether a set of replacement terms makes sense as a query reformulation, we again employ time-specific term contexts. Only, if replacement terms tend to be contained in each other's time-specific term context and thus co-occur frequently, we utilize the identified set of replacement terms as a query reformulation.

Temporal Mining

Although time-travel text search is a useful tool for accessing evolving text collections, the user is still burdened with the difficult task of digesting the contents of the potentially large set of document versions obtained in responses. In this work, we propose a framework for performing *text analytics* over such dynamically obtained subsets of the collection, in a versatile, efficient, and scalable manner. While much of the prior work has emphasized mining keywords or tags in blogs or social-tagging communities, we emphasize the analysis of *interesting phrases*. These include named entities, quotations, market slogans, and other multi-word phrases that are prominent in the dynamically derived ad-hoc subset of the corpus e.g., being frequent in the subset but relatively infrequent in the overall corpus. A particular relevant case is the comparative analysis of two ad-hoc collections against each other. The ad-hoc subsets can typically be derived as response to a keyword, time-travel, or time-period selection query.

For most practical definitions of interestingness, finding the interesting phrases requires computing the frequencies of all phrases in one or two ad-hoc subsets $\mathcal{D}'$ and $\mathcal{D}''$ of the overall corpus $\mathcal{D}$. A possible way to do this, is to scan all documents in $\mathcal{D}'$ (and $\mathcal{D}''$) using a sliding window, and for each phrase we encounter, look it up in a hash table and increase a counter maintained for it. This would be very inefficient as it requires scanning a possibly large number of documents. In [10], an *inverted phrase index* is developed to solve this problem. For each phrase, the document ids that contain it are collected into an index list, built in an IR-style inverted-file fashion [11]. In order to compute the frequencies of the phrases in $\mathcal{D}'$, all inverted lists are accessed and intersected with $\mathcal{D}'$. An approximate counting technique that intersects only a sampled subset of each list with $\mathcal{D}'$ is proposed; still, a very large number of lists has to be accessed – potentially as large as the number of phrases, regardless of how small or large the size of $\mathcal{D}'$ is.

In [1], we develop an efficient alternative to [10] with much better scalability. Documents in the entire corpus, $\mathcal{D}$, are pre-processed to extract all phrases that occur more than a specified minimum-threshold times in the corpus. These extracted phrases are then encoded and indexed to form a *phrasal signature* for each document. This approach is a major departure from the standard inverted-files paradigm that has become very common in IR and text analytics domains. Our approach resembles the notion of "forward indexes" [6, 9].

Given a subset $\mathcal{D}' \subset \mathcal{D}$ (or two sets $\mathcal{D}'$ and $\mathcal{D}''$), frequencies of phrases there needed to compute the interestingness are determined by scanning and merging the phrasal signatures of the documents in $\mathcal{D}'$ (and $\mathcal{D}''$). By using different ways of ordering and compressing the phrases in document-specific lists, different alternatives to perform the phrase mining are possible. Each of these alternatives has different capabilities in terms of reducing the search space.

References

- [1] S. Bedathur, K. Berberich, J. Dittrich, N. Mamoulis, and G. Weikum. Scalable phrase mining for ad-hoc text analytics. Research Report MPI-I-2009-5-006, Max-Planck-Institut for Informatics, Saarbrücken, Germany, 2009.
- [2] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. FluxCapacitor: Efficient time-travel text search. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds.,

33rd International Conference on Very Large Databases (VLDB 2007), Vienna, Austria, 2007, pp. 1414–1417. ACM.

- [3] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, eds., *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, Amsterdam, Netherlands, 2007, pp. 519–526. ACM.
- [4] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. Research Report MPII-I-2007-5-02, Max-Planck-Institut for Informatics, Saarbrücken, Germany, 2007.
- [5] K. Berberich, S. Bedathur, and G. Weikum. Efficient time-travel on versioned text collections. In A. Kemper, H. Schöning, T. Rose, M. Jarke, T. Seidl, C. Quix, and C. Brochhaus, eds., *Datenbanksysteme in Business, Technologie und Web (12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme")*, Aachen, Germany, 2007, *Lecture Notes in Informatics*, vol. 103, pp. 44–63. Gesellschaft für Informatik.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[7] S. Guha, K. Shim, and J. Woo. REHIST: Relative error histogram construction algorithms. In *International Conference on Very Large Databases (VLDB)*, 2004.

[8] E. J. Keogh, S. Chu, D. Hart, and M. J. Pazzani. An online algorithm for segmenting time series. In *IEEE International Conference on Data Mining (ICDM)*, 2001.

[9] L. Lim, M. Wang, S. Padmanabhan, J. S. Vitter, and R. Agarwal. Dynamic maintenance of Web indexes using landmarks. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, 2003, pp. 102–111. ACM.

[10] A. Simitsis, A. Baid, Y. Sismanis, and B. Reinwald. Multidimensional content exploration. *PVLDB*, 1(1):660–671, 2008.

[11] J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Comput. Surv.*, 38(2):6, 2006.

31.5.2 Text Classification

Investigators: Georgiana Ifrim and Gerhard Weikum in cooperation with Gökhan H. Bakir

Text categorization (also known as *text classification*) is the process of organizing text documents into predefined categories according to the documents' content. More explicitly, given a set of example documents for each category, we want to *automatically* learn the characteristics of each category, or learn a model of each category, which we can then apply to assign categories to a new (unlabeled) document. The models of each category learned this way are also called *classifiers*.

Until the late 1980's, classifiers were built manually by a set of expert users, who needed to invest a huge amount of time into finding a comprehensive set of rules for characterizing the given set of categories. Each set of rules per category would then be used to decide whether the respective category should be assigned to an unlabeled document. With the sudden increase in the amount of digital information in the late 1990's, this type of manual system building became infeasible, since manually designing application-dependent rules is both slow and very expensive. The new approaches that took over around this time were based on *machine learning techniques* and relied on automatically learning characteristics of

each category based on a set of labeled example documents. Nevertheless, many applications involve learning a supervised classifier from very few explicitly labeled training examples, since the cost of manually labeling the training data is often prohibitively high. For instance, we expect a good classifier to learn our interests from a few example books or movies we like, and recommend similar ones in the future, or we expect a search engine to give more personalized search results based on whatever little it learned about our past queries and clicked documents. There is thus a need for classification techniques capable of learning from sparse labeled data, by exploiting additional information about the classification task at hand (e.g., background knowledge) or by employing more sophisticated features (e.g., n-gram sequences, trees, graphs). In our work, we focus on two approaches for overcoming the bottleneck of sparse labeled data.

Inductive/Transductive Latent Model (ILM/TLM)

We first propose the *Inductive/Transductive Latent Model* (ILM/TLM) [1, 2, 3], which is a new generative model for text documents. ILM/TLM has various building blocks designed to facilitate the integration of background knowledge (related to the domain of the labeled data) into the process of learning from small training data. We advocate using external ontologies for instantiating the structure of our latent model, rather than selecting an appropriate structure by time-consuming model selection strategies. Such ontologies are currently growing and are freely available. We give an Expectation-Maximization algorithm for learning the parameters of ILM/TLM. The parameter space can be huge, but we propose pruning it by learning a prior on the model parameters based on background knowledge. For example, if the training data is too sparse for learning robust parameter values, a prior which relies on context-similarity in the given corpus and external sources, can help improve the final predictions by 10%. Additionally, background knowledge, e.g., encyclopedia such as Wikipedia, can be used to explicitly or implicitly extend the topic descriptions provided by the training set. Empirical results show that the additional flexibility of ILM/TLM offered by its various building blocks results in improved classification results for small training sets, as compared to other state-of-the-art classifiers. Additionally ILM/TLM has the advantage of interpretability and robustness to training/test distribution shifts. We analyze different ways of setting parameters, and the effect of each building block on the overall model performance. Last but not least, we show that taking advantage of unlabeled data, which is abundantly available in many applications, improves the results of our model.

Structured Logistic Regression (SLR)

Second, we propose *Structured Logistic Regression* (SLR) [1, 4], which is a new coordinate-wise gradient ascent technique for learning logistic regression in the space of all (word or character) sequences in the training data. SLR exploits the inherent structure of the n-gram feature space in order to automatically provide a compact set of highly discriminative n-gram features. We give theoretical bounds which quantify the "goodness" of the gradient for each n-gram candidate given its *length-(n-1)* prefix. We show that by using the proposed bounds, we can efficiently work with variable-length n-gram features both at the word-level and the character-level. Our detailed experimental study shows that while SLR achieves similar

classification results to those of the state-of-the-art methods (which use all n-gram features given explicitly), it is more than an order of magnitude faster than its opponents. We also consider the problem of learning the tokenization of the input text, rather than explicitly fixing it in advance (as in the bag-of-words model). We show that SLR can be used to learn arbitrarily sized discriminative n-grams, rather than n-grams that are restricted to a hypothesized "good" size.

Given the flexibility of SLR for learning variable-length n-gram patterns, this model could be applied to supervised information extraction for learning patterns that are indicative of binary relations. Additionally, SLR can be applied to other domains such as biological sequence classification, where mining variable-length sequences is of particular importance. Theoretical results similar to those presented in our work for learning with sequences apply directly to trees or graph representations (for example for XML documents), with only a few implementation modifications. This is true because the simple monotonicity property needed by our proofs holds also in the case of more complex structures such as trees and graphs.

The techniques developed in our work can be used to advance the technologies for automatically and efficiently building large training sets, therefore reducing the need for spending human computation on this task.

References

- [1] G. Ifrim. *A Statistical Learning Approach to Concept-Based Document Classification*. Phd thesis, Universität des Saarlandes, 2009.

[2] G. Ifrim, M. Theobald, and G. Weikum. Learning word-to-concept mappings for automatic text classification. In L. De Raedt and S. Wrobel, eds., *Proceedings of the 22nd International Conference on Machine Learning - Learning in Web Search (LWS 2005)*, Bonn, Germany, 2005, pp. 18–26.

[3] G. Ifrim and G. Weikum. Transductive learning for text classification using explicit knowledge models. In J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, eds., *PKDD 2006: 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Berlin, Germany, 2006, *LNAI 4213*, pp. 223–234. Springer. Acceptance ratio 1:7.

- [4] G. Ifrim and G. Weikum. Fast logistic regression for text categorization with variable-length n-grams. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2008)*, Las Vegas, Nevada, USA, 2008, pp. 354–362. ACM.

31.5.3 Prediction in Heterogeneous Networks

Investigators: Ralitsa Angelova, Gjergji Kasneci, and Srikanta Bedathur

Heterogeneous graphs are developing faster than ever. In the context of rapidly growing social networks (e.g. *Flickr*, *Deli.cio.us*, *LibraryThing*, *LinkedIn*), heterogeneous graphs are formed by encoding users, their postings like photos, bookmarks, book descriptions, ratings, etc., and other contextual information as graph nodes, belonging to different node types but co-existing and mutually influencing each other in the graph. Heterogeneous graphs are formed in many other areas like medical domains (containing information about patients, treatments, diseases, contacts) or e-commerce platforms (representing the complex interactions between different types of nodes like users, products, customers, rating, reviews, etc.).

We address the problem of *multi-label classification* in heterogeneous graphs, where nodes belong to different types and different types have different sets of classification labels. We present a novel approach that aims to classify nodes based on their neighborhoods. Let us use a toy example to illustrate the goal of the algorithm. If you consider the social network *Flickr*, a user is more likely to belong to the class *Outdoor activity fans* if she has frequently used tags belonging to the class *Mountain* and/or she has viewed, commented and posted photos classified as *Nature Photography*. We model the mutual influence of nodes as a random walk in which the random surfer aims at distributing class labels to nodes while walking through the graph. When viewing class labels as "colors", the random surfer is essentially spraying different node types with different color palettes; hence the name GRAFFITI of our method. In contrast to previous work on topic-based random surfer models, our approach captures and exploits the mutual influence of nodes of the same type based on their connections to nodes of other types. We base the strength of the mutual influence between nodes sharing the same type, $v, v' \in V_k$, on two criteria: the bigger the overlap of shared neighbors, the higher the influence of v and v' on each other; and the bigger the mutual influence among shared neighbors, the higher the influence of v and v' on each other.

These criteria are incorporated in the random walk the GRAFFITI surfer takes by giving him at each node the possibility to either perform a two-hop walk, spreading colors to same-type nodes via heterogeneous common neighbors - $F2$, perform a one-hop walk to any node of v's heterogeneous neighborhood – $F1$, or choose at random any node from G – perform a random jump J. All these actions are taken with the respective probabilities $P[v', v, c_i, c_i, F2]$, $P[u, v, c_j, c_i, F1$, and $P[v, c_i, J]$. More formally, the importance of a node $v \in V_k$ with respect to class c_i is given by

$$\begin{aligned}\mu_i(v) \;=\; & P[v, c_i, J] \;+\; \sum_{v' \in V_k} P[v', v, c_i, c_i, F2_{same\ color}] \\ & +\; \sum_{v' \in V_k} \sum_{j} P[v', v, c_j, c_i, F2_{change\ color}] \;+\; \sum_{j} P[u, v, c_j, c_i, F1]\end{aligned}$$

We show important properties of our algorithm such as convergence and uniqueness of the solution. We also confirm the practical viability of GRAFFITI by an experimental study on subsets of the popular social networks *Flickr* and *LibraryThing*. We demonstrate the superiority of our approach by comparing it to three other state-of-the-art techniques for graph-based classification [1, 2, 4, 5].

Another significant problem that we address is the *prediction of links* in a social network representing interactions between individuals – both new as well as recurring ones. Adequate link predictions build the core of successful product recommendations and targeted advertising, or new friendship suggestions in a social network. Since the seminal work of Liben-Nowell and Kleinberg [3], there has been a significant interest in considering only the graph structure to make link predictions. However, most proposed methods consider a single snapshot of the network as the input, neglecting an important aspect of these social networks, namely, their evolution over time. In our work [6], we investigate the value of incorporating the history information available on the interactions (or links) of the current social network state. Our results show that considering time-stamps of past interactions significantly improve the

prediction accuracy of new and recurrent links over rather sophisticated methods proposed recently [7]. Further, we introduce a novel testing method necessary to evaluate the benefits of incorporating time awareness in the link prediction.

References

- [1] R. Angelova, G. Kasneci, F. M. Suchanek, and G. Weikum. GRAFFITI: Node labeling in heterogeneous networks (poster). In *18th International World Wide Web Conference (WWW 2009)*, Madrid, Spain, 2009. ACM.

[2] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[3] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, New York, NY, USA, 2003, pp. 556–559. ACM.

[4] L. Nie, B. D. Davison, and X. Qi. Topical link analysis for Web search. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2006, pp. 91–98. ACM.

[5] H.-J. Oh, S. H. Myaeng, and M.-H. Lee. A practical hypertext catergorization method using links and incrementally available class information. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2000, pp. 264–271. ACM.

- [6] T. Tylenda. Time-aware link prediction in evolving social networks. Masters thesis, Universität des Saarlandes, 2009.

[7] C. Wang, V. Satuluri, and S. Parthasarathy. Local probabilistic models for link prediction. In *ICDM*, 2007, pp. 322–331. IEEE Computer Society.

31.5.4 Learning from Preferences

Investigators: Bilyana Taneva and Gerhard Weikum

In many applications we are given a set of objects and a set of preference relations between them. For example, consider the set of retrieved documents from a query, where the users express their preferences by determining some documents as relevant and others as non-relevant (e.g., by clicking or not clicking on a link in the top-10 result summaries). Another possible application are scheduling tasks, where the objects are time slots and the preferences represent the availability of the different individuals. The goal is, given the training preference data, to learn a ranking function that ranks all object accordingly.

The problem of learning from preference data has two basic ways of interpretation. It can be considered as a problem of finding an objective ranking function based on subjective preferences or it can be solved only with respect to individual persons. The interpretation depends on the specific application, so can, for example, the personalized view be used to improve the retrieval quality of search engines by keeping records of users' click preferences.

Conjoint analysis [1] is a family of techniques popular in market research which is highly related to the problem of learning to rank from preferences. The main tasks of conjoint analysis are to assess and measure an individual's or a population's preferences over a set of objects that can be described by parameters and their levels. In our work [3], we deal with

choice based conjoint analysis studies where one observes individual's choices over a small subset of object options. We analyze the user preferences using two approaches, one based on statistical assumptions, and one based on a direct regression approach.

There are many other ways to obtain preference relations between objects. Some of them are to collect explicit feedback when asking people to order different objects, or to derive implicit feedback when considering, for example, user clicks on documents. In our research we investigate further possibilities for obtaining and analyzing the preference relations. Currently we are investigating the ranking of images for given entities (e.g., famous people, landmarks) that have some related facts and properties. We use the facts corresponding to a specific entity to pose a sequence of queries to search engines, based on the overlap of the retrieved images for the different queries we derive preferences between the images. Finally, using the Ranking SVM algorithm [2] or some other optimal regression algorithm, we find the ordering of the images for the corresponding entities.

Another direction to explore is the problem of learning from preferences over more complicated data that has some underlying structure. The objects in such data could be represented as graph nodes and the edges of the graph could represent relationships between the corresponding objects. In this case, preferences for some objects influence other preference choices. The additional structure could thus improve the quality of the rankings.

References

[1] A. Gustafsson, A. Herrmann, and F. Huber. Conjoint analysis as an instrument of market research practice. In A. Gustafsson, A. Herrmann, and F. Huber, eds., *Conjoint Measurement. Methods and Applications*, 2000, pp. 5–45. Springer, Berlin.

[2] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2002, pp. 133–142.

- [3] B. Taneva, J. Giesen, P. Zolliker, and K. Mueller. Choice based conjoint analysis: Discrete choice models vs. direct regression. In E. Hüllermeier and J. Fürnkranz, eds., *Proceedings of the ECML/PKDD Workshop on Preference Learning (PL-08)*, Antwerp, Belgium, 2008, pp. 67–81.

31.5.5 XML Summarization

Investigators: Maya Ramanath, Kondreddi Sarath Kumar, and Georgiana Ifrim

With the ubiquity of XML as the format of storage and exchange of data, we can expect to see ever-growing repositories of XML documents. Exploration of these collections requires the use of a diverse set of tools ranging from classifiers, clustering tools, data visualizers to mining software. One of the ways in which *human-centric* exploration can be made easier is to provide the user with a concise, summarized view of the information contained in an individual XML document or an entire set of documents. Consider the following scenario. Suppose there is a large corpus of XML documents, each of which describes a movie released in the last 30 years, for example, extracted from IMDB (`http://www.imdb.com/`). A movie enthusiast wants to make a list of interesting movies based on various criteria such as genre, lead actors, directors, etc. She first decides to narrow the focus to just thrillers. However, she then has to look into each document individually, since only then is it possible for her to tell whether the combination of actors, directors, etc. interests her. This would be time-consuming

if the documents in question contain hundreds of tags each. Instead, if short summaries of each document could be presented to her, she could use these summaries to filter out movies she certainly would not be interested in. The generation of such summaries is the problem we address in this work.

We regard the problem of generating XML summaries as a two-stage problem. First, we separately rank tags and text according to a notion of their importance. Next, we construct the summary based on the tag and text scores. The choice of tag-text pairs is made such that the summary reflects only their relative importance in the document, thus achieving a balance between including importance and coverage.

Ranking Tags. We use two criteria for considering a tag t important. First, the typicality of tag t is its salience in the corpus. For example, `title` is the most salient tag in the corpus, since it is present in all documents and is important because it sets the context for the rest of the document. Second, the specialty of a tag is a measure of how important the tag is in the current document. For example, `production_location` may occur once typically, but if the current document has 10 of those, then it implies that the film was shot in an unusually large number of locations. Also, if the current document contains `oscar_winner` while the average document does not, then that too should be considered special.

Taking these two aspects into consideration, we use the following to compute the probability of a tag:

$$P(T_i) = \alpha P_{typ}(T_i) + (1 - \alpha) P_{spe}(T_i)$$

where $P(T_i)$ is the probability of choosing tag T_i, $P_{typ}(T_i)$ and $P_{spe}(T_i)$ are the probabilities of choosing T_i based on its typicality and specialty respectively, the parameter α, $0 \leq \alpha \leq 1$ is set by the user or learned through examples.

Typicality is calculated as the fraction of corpus documents containing tag T_i while specialty is calculated as the deviation in the number of T_i's given an "average" document in the corpus.

Ranking Text. We divide text into the following categories: i) entities and ii) regular text. Entities are treated holistically and our system currently supports proper names. Regular text could be long or short and can be reduced to a set of terms. Text values are ranked based on their occurrence in three different contexts: the tag context, the document and the corpus and their probability is calculated as:

$$P(t_j|D, T_i) = \lambda P(t_j|D, c(T_i)) + \mu P(t_j|D) + (1 - \mu - \lambda) P(t_j|C)$$

where the first term, $P(t_j|D, c(T_i))$ denotes the probability of choosing t_j within the *context* of T_i (denoted $c(T_i)$). The second and third terms, $P(t_j|D)$ and $P(t_j|C)$ denote the probability of t_j in the document and the corpus respectively.

Generating the Summary. Given a list of tags and text values ordered by their probabilities, as well as a memory budget (in terms of how many tag-text pairs can be included) we now need to select the best summary which contains important and diverse tags. We do this in

two steps. First, the tags are selected in decreasing order of probabilities and in proportion to their probabilities. For example, if we had 2 tags `A` and `B` with probabilities 0.7 and 0.3, then to construct a summary of 10 elements, we would choose 7 `A`'s and 3 `B`'s. Once the tag structure is fixed, we now choose the best possible text values for each tag. In our previous example, we choose the 7 best text values for `A` and the 3 best text values for `B`.

Evaluation

We conducted extensive user evaluations on movie and people datasets derived from IMDB. Summaries of different sizes and with different parameters were evaluated by at least 3 evaluators. In addition, a gold-standard summary was created for each size and each document and was added to the evaluation set. In total 80 automatically generated summaries were evaluated. Over 75% of the summaries were deemed to be acceptable by at least 2 evaluators.

Xoom: Zooming in and out of XML documents

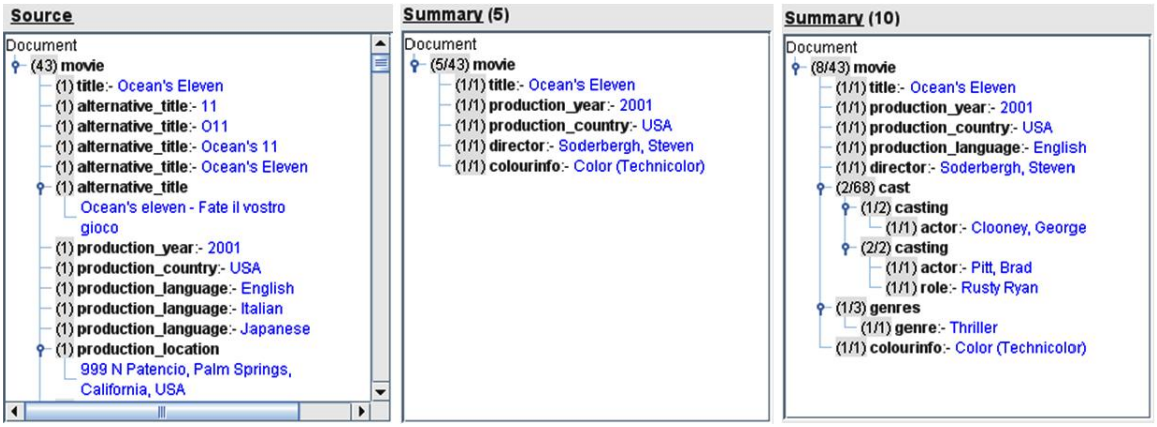


Figure 31.3: "Ocean's Eleven" – Excerpt from the original document and its 5-element and 10-element summaries

We built a tool to explore XML documents using the techniques described above. The tool has two parts. The first part allows users to "zoom-out" of the XML document. That is, the specified XML document and the desired summary size are taken as input and a summary of the given size is generated and presented to the user. Figure 31.3 shows the 5-element and 10-element summaries of the movie Ocean's Eleven. The original document is shown on the left-hand side of the screen. Though not shown in the figure, users may provide the desired size of the summary in a separate dialog. In both the summary snapshots in Figure 31.3, the numbers in the parentheses before each tag indicate *# of children displayed / total # of children* for that tag.

The second part of the tool allows users to "zoom-in" to specific parts of the XML documents. For example: "top-10 actors in the movie", or "top-5 tags in this document". The user chooses the XML document to zoom into, as well as the elements (tags or text values)

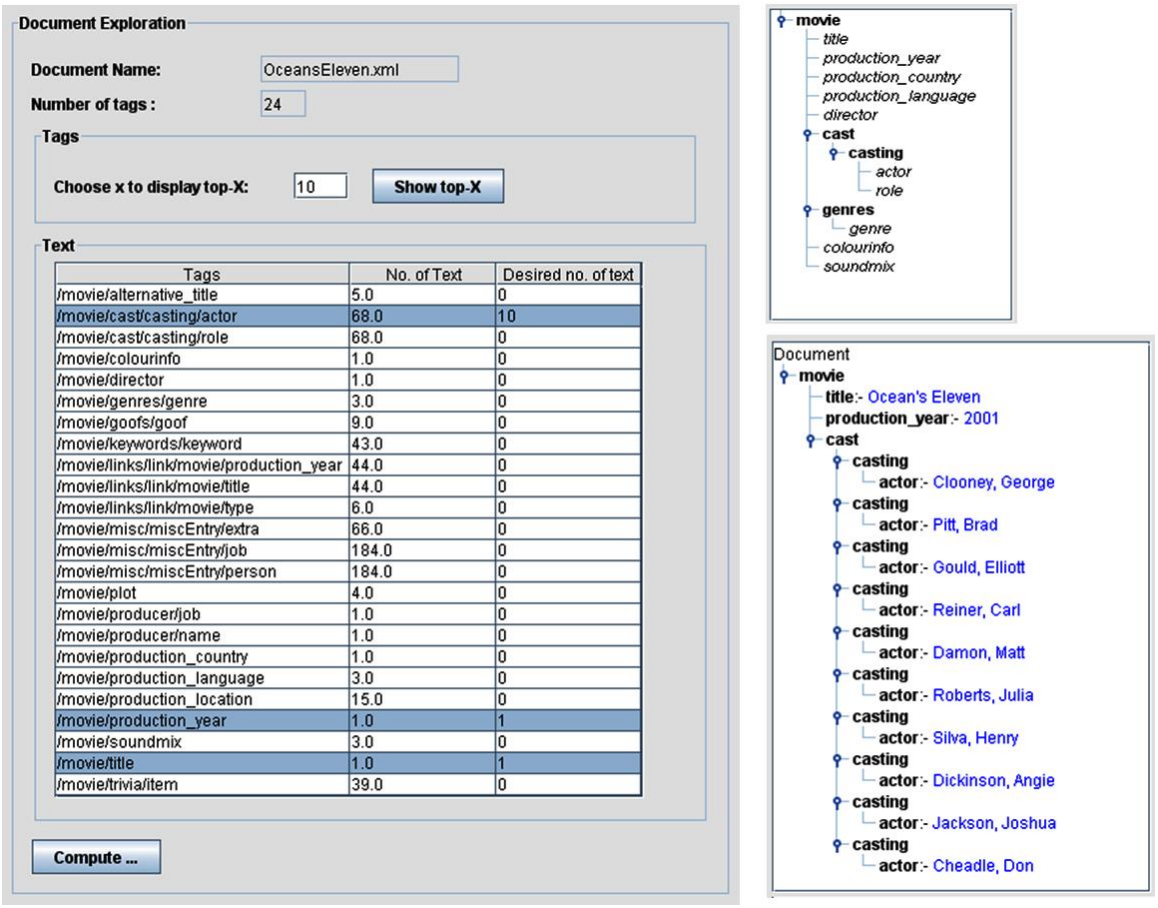


Figure 31.4: Zooming in: Generating the top-k tags and text values

and the number of elements she wants to see. Figure 31.4 shows an example. The dialog on the right hand side has two parts – one for tags and one for text. In the first part for top-k tags, users can specify how many ranked tags she would like to see. The second part for text shows a table with three columns. The first and second columns show the paths available in the document (or the summary, if the user wants to zoom in based on the tags in the current summary) and the total number of text values available for each path respectively. In the third column, users can specify how many text values for each tag should be shown. The results of choosing to zoom in to the top-10 tags (out of a possible 24) and the top-10 actors (out of a possible 68) are shown in the right hand side of the figure (note that these are two independent functions).

Summary of Contributions

Though the field of text summarization has been around for several years now [1], it is not directly applicable to XML summarization due to both, the presence of structure (tags) as well as the nature of text values (short text such as names, in addition to long text values).

Our main contribution is a set of techniques to *automatically* generate a summary of given size for an XML document. We make the assumption that the document is part of a larger corpus of similar documents. We collect and use statistics from both the document as well as the corpus to determine the important tags and text values that should go into the summary. Our work has been published in [2] and a demonstration of a tool implementing the above ideas has been accepted [3].

References

[1] I. Mani. Summarization evaluation: An overview. In *Proc. of NTCIR Workshop*, 2001.

- [2] M. Ramanath and K. Sarath Kumar. A rank-rewrite framework for summarizing XML documents. In *2nd International Workshop on Ranking in Databases (DBRank), ICDE 2008 Workshops*, Cancun, Mexico, 2008, pp. 540–547. IEEE Computer Society.
- [3] M. Ramanath and K. Sarath Kumar. Xoom: A tool for zooming in and out of XML documents. In *Proceedings of the 12th International Conference on Extending Database Technology (EDBT 2009)*, St. Petersburg, Russia, 2009.

31.5.6 Web Archive Quality

Investigators: Marc Spaniol, Dimitar Denev, and Arturas Mazeika

Despite initiatives of the International Internet Preservation Consortium (IIPC) (`http://netpreserve.org/`) and the Internet Archive (`http://www.archive.org/`) in capturing Web contents for future generations, limitations such as storage space, bandwidth, and crawling politeness as well as threats such as Web spam and crawler traps heavily affect the crawling performance and, thus, the quality of the collected data. Current methods are based on snapshot crawls and "exact duplicate" detection [1]. The coherence of data in terms of proper dating and proper cross-linkage is influenced by the temporal limitations (long duration, low recrawling frequency, etc.) of the crawl process.

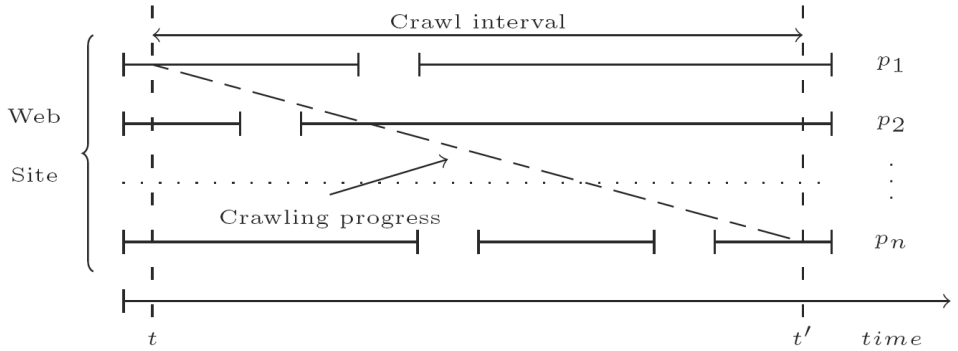


Figure 31.5: Timeline of a Web site crawling process

Web archiving is commonly understood as a continuous process that aims at archiving the entire Web (broad scope). However, a typical scenario in archiving institutions or companies

is to periodically—e.g. monthly—create high quality captures of a certain Web site. These periodic domain-scope crawls of Web sites aim at obtaining a best possible representation of a site. A reason for customers having their site archived on a regular basis is, for instance, to guard themselves against accusations regarding intellectual property rights, fraud or non-compliance with legal requirements (e.g. EU laws about imprints, terms of use, etc.). Figure 31.5 contains an abstract representation of such a domain-scope crawling process. This Web site consists of n pages $(p_1, \ldots, p_n)$. Each of them consists of several successive versions, indicated by the horizontal lines (e.g., p_n has three different versions in $[t; t']$). Ideally, the result of a crawl would be a complete and instantaneous snapshot of all pages at a given point of time. In reality, one crawl requires an extended time period to gather all pages of a site while being potentially modified in parallel, thus causing incoherencies in the archive. The risk of incoherence increases further due to politeness constraints (i.e., pausing a few seconds between successive HTTP requests) and need for sophisticated time-stamping mechanisms.

An ideal approach to Web archiving would be to have captures for every domain at any point in time whenever there is a (small) change in any of the domain's pages. Of course, this is absolutely infeasible given the enormous size of the Web, high content-production rates in blogs and other Web 2.0 venues, the disk and server costs of a Web archive, and also the politeness rules that Web sites impose on crawlers. We therefore settle for the realistic goal capturing Web sites at convenient points (whenever the crawler decides to devote resources to the site and the site does not seem to be highly loaded), but when doing so, the capture should be as "authentic" as possible. In order to ensure an "as of time point x (or interval $[x, y]$)" capture of a Web site, we develop an archiving crawler that *ensures coherence* of crawls regarding a time point or interval, and *identifies* those contents of the Web site that *violate coherence*.

We have developed a coherence framework that is capable of dealing with properly as well as improperly dated contents (where dating refers to HTTP last-modified timestamps) [2]. Depending on the data quality provided by the Web server, we have developed different coherence optimizing crawling strategies, which outperform existing approaches and have been tested under real life conditions. Moreover, by using a smart revisit strategy for crawlers, we are also capable of discovering and (as a consequence) ensuring coherence for contents, which are improperly dated and not correctly interpretable with conventional archiving technologies. This way, we improve the quality of Web archives, regardless of how unreliable Web servers are. We can layout a site capture's spanning tree and visualize its coherence defects by applying graphML compliant software. This visual metaphor is intended as an additional means to automated statistics for understanding the problems that occurred during capturing. Figure 31.6 depicts a sample visualization of an `mpi-inf.mpg.de` domain capture (about 65.000 pages) with the Visone software (`http://visone.info/`). Depending on the nodes' sizes, shapes, and colors the user gets an immediate overview on the success or failure of the capturing process. In particular, a node's size is proportional to the amount of coherent web contents contained in its sub-tree. In the same sense, a node's color highlights its "coherence status". While green stands for coherence, the signal colors yellow and red indicated (content modifications and/or link structure changes). The most serious defect class of missing contents is colored in black. Finally, a node's shape indicates its MIME type ranging from circles (HTML contents), hexagons (multimedia contents), rounded rectangles (Flash or similar), squares (PDF contents and other binaries) to triangles (DNS lookups).

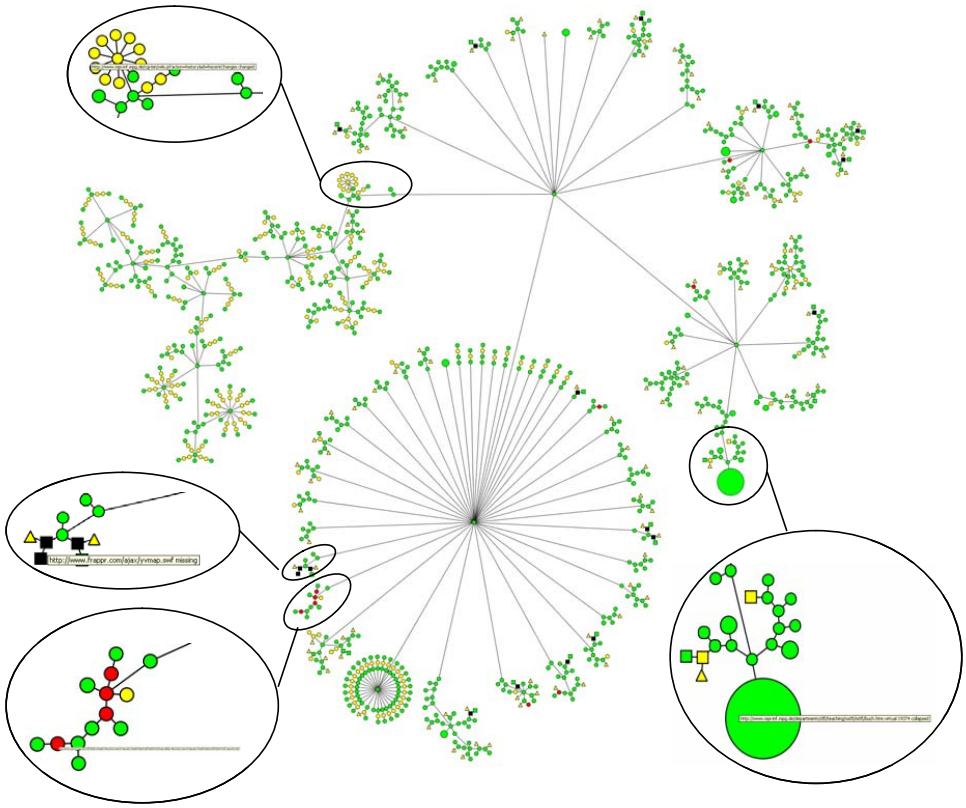


Figure 31.6: Coherence defect visualization

This visualization helps the crawl engineer to better understand the nature of change(s) within Web sites and adapt the crawling strategy for future captures.

References

[1] J. Masanès. *Web Archiving.* Springer, New York, Inc., Secaucus, NJ, 2006.

- [2] M. Spaniol, D. Denev, A. Mazeika, and G. Weikum. Data quality in web archiving. In *Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW 2009) in conjunction with the 18th World Wide Web Conference (WWW 2009)*, Madrid, Spain, 2009, pp. 19–26. ACM.

31.5.7 Visual Data Mining

Investigators: Arturas Mazeika in cooperation with Andrej Taliun and Michael H. Böhlen

Visual Analytics of Hierarchical Heavy Hitters

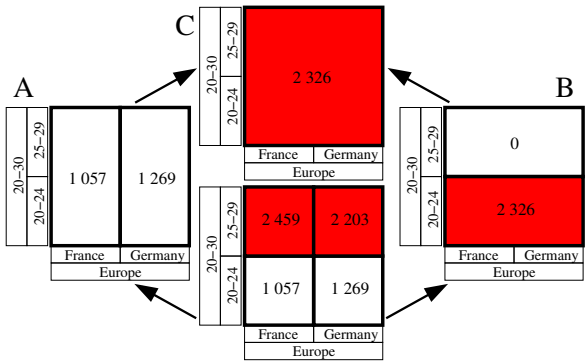
Data is being produced at a much faster rate than our ability to analyze it. Analysts and decision makers turn to visual analytics to understand their data, obtain overviews, drill down into interesting aspects, and derive insights. Visual analytics represents the information visually, allowing the human to directly interact with the visualization.

In our work [1], we focus on so-called Visual Hierarchical Heavy Hitters (VHHHs) [5] and develop a methodology for how to analyze and interpret the underlying data. Related work in this area focus on efficiency aspects in static data and computation of Hierarchical Heavy Hitters (HHH) for streaming data. In contrast, our work focuses on a methodology on how to analyze the data with VHHH, in order to understand the capabilities and limits of both of the VHHH method and the HHH as a data summary.

HHHs are defined over categorical, multi-dimensional data such as biological taxonomies or census data. With each dimension, we associate a dimension hierarchy such as the hierarchy of all species in the evolution in biology or a division of census data into administrative areas. Dimension hierarchies in multi-dimensional datasets generate a lattice structure (cf. Figure 31.7(b)). All tuples exceeding a given threshold are called HHHs (e.g., age 25–29 in France with count 2,459 and age group 25–29 in Germany with count 2,203 in Figure 31.7(b)). The remaining non-HHH counts are grouped along the lattice and further compared to the threshold.

(**Location**, **Age**)	**Count**
(Europe/France, 20-30/20-24)	1,057
(Europe/France, 20-30/25-29)	2,459
(Europe/Germany, 20-30/20-24)	1,269
(Europe/Germany, 20-30/25-29)	2,203

(a) Example Dataset



(b) HHHs, The Lattice Structure

Figure 31.7: HHHs

HHHs are used to identify "big players" in the dataset. All HHHs would be returned in the answer table even if there are hundreds or thousands of HHHs for a given query. Our work shows that HHHs are a powerful visual data mining tool and can be successfully used to analyze the dependencies between the dimensions, as well as for reasoning about the distribution of the data. In our analysis, we experimented with three real world datasets captured as binary relations: *immigration*, *medals*, and *firewall*. The *immigration* datasets records the country of origin and the age of the immigrants to a country (e.g., *(Europe/France, 20-30/20-24, 1,057)*), *medals* records the country and sport type of athletes that won gold medals (e.g., *(World/Europe/EasternEurope/Russia, Olympics/Summer/Outdoor/Athletics, 2)*), while *firewall* records information about programs accessing Internet sites (e.g., *(Programs/Office/Access, com/microsoft/db, 100)*). We identified four patterns: flag, chess-board, cross, and triangular to describe the dependencies between the dimensions and distributions of the datasets. We describe them below.

Figures 31.8(a)–31.8(b) illustrate the *flag pattern*. In the pattern either all or none of the nodes are HHHs along one dimension (cf. the vertical axis in the figures) and HHH nodes are followed by non HHH nodes along the other dimension (cf. the horizontal axis). This

shows a steady behavior for one dimension and periodicity for the other dimension. The flag pattern is prominently seen in the *immigration* dataset in Figure 31.8(b). Their four stripes of HHHs correspond four continents of the world (cf. the x-axis in the figure). In two continents there are two countries that dominate the others (cf. very thin and tall rectangles in the figure), one continent has only young people immigrants and one continent has either young or retired immigrants.

In the *chess board pattern* (cf. Figures 31.8(c)–31.8(d)) one or several HHH nodes are followed by one or several non-HHH nodes along both dimensions. The chess board pattern is easily seen in the *medals* dataset in Figure 31.8(d). The figure shows that there is a group of countries that established a strong track record in certain areas of sports with individual winners varying from country to country and from one sport to the other.

The *cross pattern* (cf. Figure 31.8(e)–31.8(f)) consists of two HHHs crossing each other. This indicates a very scattered dataset consisting of a wide variety of tuples with very small values. Figure 31.8(f) illustrates the cross pattern on the *firewall* dataset. The vertical stripe corresponds to the Internet browser accessing multiple sites, while the horizontal stripe shows the accesses by numerous programs of the same resource on the server.

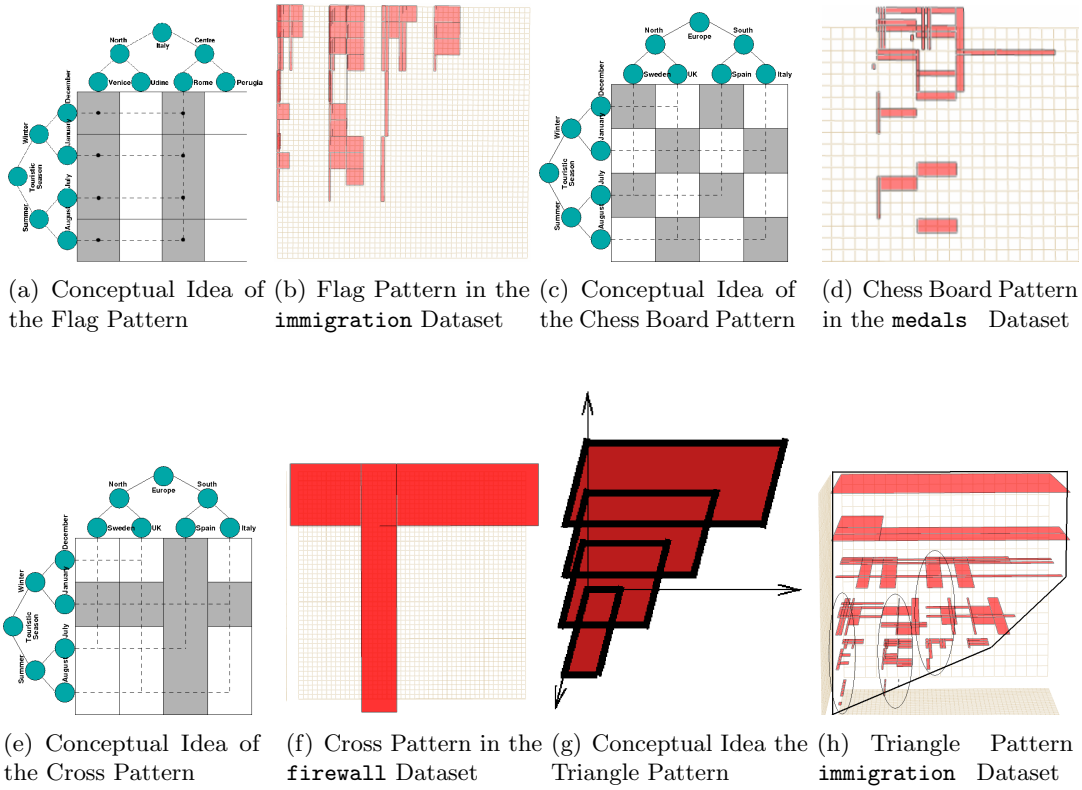
In the *triangular pattern* (cf. Figures 31.8(g)–31.8(h)) layers of HHHs form a triangle or multiple triangles . The triangle distribution of layers of HHHs indicates a skew (Zipfian distribution) in the data. Typically, in such datasets there is a dominant tuple (lowest level of grouping), a dominant group (1^{st} level of grouping), a dominant group of groups (2^{nd} level of grouping), and so on.

CORE: Non-parametric Clustering of Large Numeric Databases

Moreover, we have developed a clustering method coined CORE [2] to group numeric multi-dimensional data in an unsupervised manner. CORE is based on explicit local maxima in the statistical density [3] of the data and makes the clustering robust to noise, identifies arbitrarily shaped clusters, does not require the user to provide the number of clusters upfront, and does not impose restrictions on the dimensionality or the closeness (or even overlap) of clusters.

The explicit computation of the local maxima of the density faces two challenges. First, the density maxima of numeric datasets are affected by small density fluctuations. Density fluctuations produce false peaks, and a robust identification of a local maximum is non-trivial. Second, the explicit computation of the maxima of the density quickly becomes prohibitively expensive for large multi-dimensional datasets. CORE successfully solves these challenges with the help of gradients of the density, rectangular neighborhoods of grid points, and a sequential organization of the density of multi-dimensional datasets.

CORE does not assume that clusters follow any a-priori model and does not require model parameters that guide the clustering process. CORE is a non-parametric method since the clustering only depends on the precision of the density estimation of the adaptive density tree (AD-Tree), a data structure that we proposed in [4]. The precision of the AD-Tree is controlled by ε, which is not a model parameter and exhibits a monotonic behavior for values larger than the kernel bandwidth: a decrease of ε yields an increase of the estimation precision of the AD-Tree and, hence, a higher quality clustering (at the cost of a higher runtime). In contrast, related clustering techniques depend on various model parameters like the number of clusters, the number of cluster representatives, a shrinking parameter, and the



(a) Conceptual Idea of the Flag Pattern

(b) Flag Pattern in the `immigration` Dataset

(c) Conceptual Idea of the Chess Board Pattern

(d) Chess Board Pattern in the `medals` Dataset

(e) Conceptual Idea of the Cross Pattern

(f) Cross Pattern in the `firewall` Dataset

(g) Conceptual Idea the Triangle Pattern

(h) Triangle Pattern `immigration` Dataset

Figure 31.8: VHHH Patterns

size of the neighborhood.

Table 31.1 compares the quality of CORE with the state-of-the-art methods CURE and OPTICS on selected synthetic datasets. We use standard precision, recall, and F-Score measures to assess the quality of the clusterings. CORE clearly outperforms the other methods for databases with complex shaped clusters (e.g., the two spirals in the first row in the table). Here, CORE correctly identifies both spirals and has the highest average F1-score. CURE and OPTICS accurately identify the inner spiral, even outperforming CORE a bit (OPTICS with a F1-score of 99.6 vs. CORE with 97.1). However, all competitors are substantially worse for the outer spiral.

The *Hierarchical Clusters* dataset is the most challenging for all clustering techniques. The dataset consists of a plane, two dense spheres embedded in the plane cluster, and a sparse sphere outside the plane. CORE performs very well, while the competitors perform poorly: CURE merges half of the plane and the embedded spheres into one cluster, OPTICS splits the plane and removes many border points from the spheres.

The *Overlapping Spheres* dataset compares techniques for two overlapping spherical clusters. All techniques correctly identify the two clusters, but CURE incorrectly clusters some of the points of the left sphere, and OPTICS removes all points where the density level is lower than the density level of the overlap.

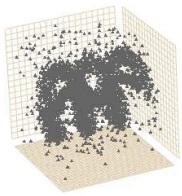
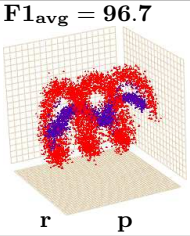
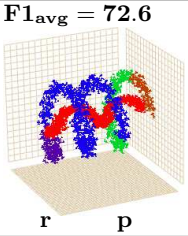
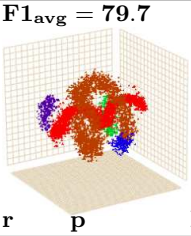
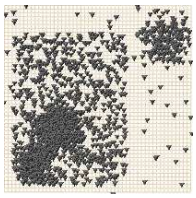
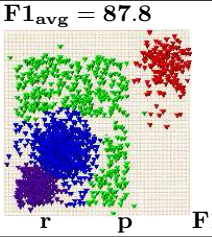
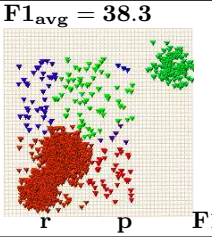
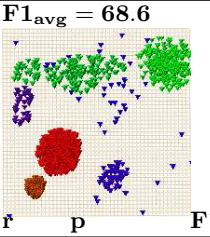
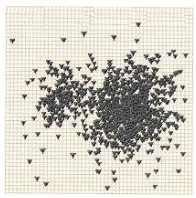
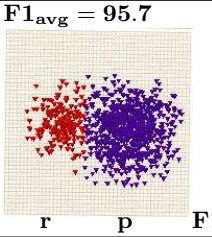
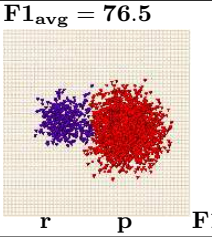
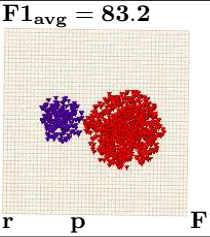
Dataset	CORE			CURE			OPTICS and DataBubbles		
Two Spirals	**$F1_{avg}$ = 96.7**			**$F1_{avg}$ = 72.6**			**$F1_{avg}$ = 79.7**		
									
	r	**p**	**F1**	**r**	**p**	**F1**	**r**	**p**	**F1**
Inner Spiral	100.0	94.27	97.1	91.8	99.9	95.7	99.7	99.5	99.6
Outer Spiral	92.88	100.0	96.3	32.9	100.0	49.5			
Hierarchical Clusters	**$F1_{avg}$ = 87.8**			**$F1_{avg}$ = 38.3**			**$F1_{avg}$ = 68.6**		
									
	r	**p**	**F1**	**r**	**p**	**F1**	**r**	**p**	**F1**
Embedded Sphere 1	99.1	86.9	92.5	0.0	0.0	0.0	95.4	88.3	91.7
Embedded Sphere 2	98.8	80.8	88.9	95.0	53.7	68.6	53.4	100.0	69.6
Outside Sphere	99.9	99.7	99.8	55.0	100.0	70.9	86.9	100.0	93.0
Plane	56.8	90.8	69.9	7.3	97.1	13.6	11.2	94.5	20.0
Overlapping Spheres	**$F1_{avg}$ = 95.7**			**$F1_{avg}$ = 76.5**			**$F1_{avg}$ = 83.2**		
									
	r	**p**	**F1**	**r**	**p**	**F1**	**r**	**p**	**F1**
Right Sphere	97.9	98.6	98.2	72.3	99.6	83.8	80.3	98.7	88.6
Left Sphere	94.4	92.0	93.2	57.6	86.6	69.2	65.9	95.0	77.8

Table 31.1: Numerical and Visual Comparison of CORE

References

- [1] A. Mazeika, M. H. Boehlen, and D. Trivellato. Analysis and interpretation of visual hierarchical heavy hitters of binary relations. In P. Atzeni, A. Caplinskas, and H. Jaakkola, eds., *12th East European Conference on Advances in Databases and Information Systems (ADBIS 2008)*, Pori, Finland, 2008, *LNCS 5207*, pp. 168–183. Springer.

- [2] A. Mazeika, A. Taliun, and M. H. Böhlen. CORE: Nonparametric clustering of large numeric databases. In *SIAM International Conference on Data Mining (SDM 2009)*, Sparks, Nevada, 2009. Society for Industrial and Applied Mathematics.

[3] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

[4] A. Taliun, M. Böhlen, and A. Mazeika. Adaptive density estimation. In *International Conference on Very Large Databases (VLDB)*, 2006, pp. 1191–1194.

[5] D. Trivellato, A. Mazeika, and M.H.Böhlen. Using 2D hierarchical heavy hitters to investigate

binary relationships. In *LNCS/LNAI State-of-the-Art Survey*, vol. 4404, pp. 215–236. Springer-Verlag, 2008.

31.6 Ranking and Uncertain Data Management

Coordinator: Martin Theobald

Ranking query results is a fundamental building block connecting database (DB) technology with information retrieval (IR) methods, the latter taking into account not only the *efficiency* but also the *effectiveness* of the query processing from a user perspective. Our work on combining DB and IR tries to incorporate a wealth of new research aspects, including semistructured data management (Subsection 31.6.1), statistical language models for top-k keyword search on data graphs (Subsections 31.6.2 and 31.6.3), as well as personalized search (Subsection 31.6.4) and investigating how relevance evolves over time (Subsection 31.6.5).

Moreover, *uncertain data management* has recently become an emerging issue in both database theory and the design of new database systems. Managing uncertainty involves aspects from logics, machine learning, and probability theory. New issues in managing uncertainty also arise with the advent of new data models, in particular non-schematic, graph-based RDF/RDFS representations. These settings go beyond purely relational data or just probabilistically enhanced relational data, where fixed schemata and strong independence assumptions among data objects are still prevalent. Consequently, for these graph-based settings, graphical probabilistic models like Bayesian Nets, Markov Networks, and Markov Logic call for an efficient database infrastructure (Subsection 31.6.6).

All these probabilistic models have one thing in common: ranking. Thus, more generally speaking, ranking can be thought of as a means of managing uncertainty about the user's information need.

31.6.1 TopX: Efficient Search and Ranking for Heterogeneous XML Data

Investigators: Martin Theobald and Ralf Schenkel

TopX is a long-running project at the Max Planck Institute for Informatics and has become an efficient and mature full-text search engine for semistructured data and XPath-like queries. Extensive experiments and our ongoing participation in the annual INEX[3] benchmark series confirm state-of-the-art query processing results in terms of both effectiveness and efficiency [4, 5, 12]. As of 2007, TopX has undergone a major refurbishing step, including a complete reimplementation of the original Java prototype in C++ [1], along with a new, file-based index management and its own internal cache management. Our current prototype, coined TopX 2.0 [12], is a major step beyond the limitations of a purely relational storage back-end, towards a more compact and efficient, object-oriented storage for text-centric XML data.

TopX has been demonstrated at the 2007 Sigmod Conference [18], and Martin Theobald has won several awards for his PhD thesis *"TopX – Efficient and Versatile Top-k Query Processing for Text, Semistructured, and Structured Data"* [11], including an ACM Sigmod

[3] *INitiative for the Evaluation of XML Retrieval*

Dissertation Award Honorable Mention. Several recent B.Sc. [10] and M.Sc. [1, 6, 7, 8] thesis have been conducted in the context of TopX. An open-source version of the Java prototype has been made available for download at `http://topx.sourceforge.net/`, and a similar release of the new C++ prototype for TopX 2.0 is intended for the near future.

A Complete Framework for Top-k-Style, Non-Conjunctive XML Full-Text Search

TopX is a native IR-engine for semistructured data with XPath-like but non-conjunctive ("andish") query evaluation, which is particular challenging for efficient evaluation of path queries because of the huge intermediate set of candidate result elements, i.e., when any XML element in the collection matching any of the query conditions may be a valid result. In this "andish" retrieval mode, the result ranking is driven by score aggregations, while the query processor needs to combine both content-related and structural aspects of content-and-structure (CAS) queries into a single score per result element. High scores for some query dimensions may compensate weak (even missing) query matches, including structural conditions, at other query dimensions. Thus, the query processor may dynamically relax query conditions if too few matches would be found in a conjunctive manner, whereas the ranking allows for the best (i.e., top-k) matches to be cut off if too many results would be found otherwise. XPath queries are more difficult to evaluate in "andish" mode than typical DB-style conjunctive queries, as we can no longer use conjunctive merge-joins (corresponding to intersections of index list objects), but we need to find efficient ways of merging index lists in a non-conjunctive manner (corresponding to unions of index list objects, or "outer-joins" in DB terminology). Top-k-style evaluation strategies are crucial in these XML-IR settings, not only for pruning index list accesses (i.e., saving disk I/O's) but also for pruning intermediate candidate objects that need to be managed dynamically (i.e., saving memory space) at query processing time. Pruning also the in-memory data structures is particularly crucial for good runtimes if CPU-time becomes a dominating factor of the query processing (e.g., for complex XPath expressions, or when many index lists are already cached).

From Relational Backend to Object-Oriented Storage

TopX 2.0 combines query processing techniques from our original TopX prototype [13, 16] and carries ideas for block-organized inverted index structures for text from our IO-Top-k algorithm [2, 3] over to the XML case. The result is a novel, object-oriented storage for text-oriented XML data with sequential, stream-like access to all index objects. Just like the original engine, TopX 2.0 can also employ more sophisticated cost models for sequential and random access scheduling along the lines of [13].

Core of the new engine is a multiple-nested block-index structure that seamlessly integrates top-k-style sequential (aka. "sorted") access to large blocks stored as inverted files on disk with in-memory merge-joins for efficient score aggregations. The main challenge in designing this new index structure was to reconcile three different paradigms in search engine design: 1) sorting blocks in descending order of the maximum element score they contain for threshold-based candidate pruning and top-k-style early termination; 2) sorting elements within each block by their id to support efficient in-memory merge-joins; and 3) encoding both structural and content-related information into a single, unified index structure [12].

Moreover, TopX 2.0 now comes with its own indexing component that makes it completely independent of any underlying relational database system. While our initial TopX 2.0 prototype had already been introduced at the INEX 2007 workshop [5]—supporting only content-only (CO) queries at that time—we meanwhile made the step to a full-fledged, highly efficient XML-IR engine for XPath Full-Text. As such, TopX 2.0 has been extended to the full functionality of the previous TopX prototype, including support for arbitrary CAS queries, probabilistic candidate pruning and random access scheduling, as well as a CPU-friendly index compression scheme, which altogether yields 20-30 times better runtimes and an order-of-magnitude in storage savings compared to the previous Java/Oracle-based architecture. Both our cost-based scheduling models and the physical index block sizes can now be tuned to fit to almost arbitrary storage and middleware characteristics. This extended adaptability will also give rise to some interesting experiments on different index access scheduling strategies and block sizes over new storage media, like, for example, flash-memory disks with a lower sequential bandwidth but much better random access throughput than conventional hard disks.

INEX 2008 Ad-Hoc and Efficiency Tracks

The new architecture has been presented at two Dagstuhl events: the Ranked XML Querying[4] seminar and the INEX 2008 workshop[5]. We also demonstrated the good performance in comparison to a number of other XML-IR systems at the new INEX Efficiency Track 2008 [12, 14], which showed a factor of up to 10 better runtimes than our best competitors, at a comparable result quality in terms of both early precision (i.e., at the top-1 to top-10 result elements) and mean-average precision (MAP) (i.e., using the top-1,500 result elements). That is, TopX performed consistently well in both the INEX 2008 Ad-Hoc [4] and Efficiency Tracks [12], the latter over a mixture of Ad-Hoc queries and high-dimensional query expansions with partly more than 100 keywords. TopX 2.0 achieved overall running times of less than 51 seconds for the entire batch of 568 Efficiency Track queries in their content-and-structure (CAS) version and less than 29 seconds for the content-only (CO) version, respectively, using a top-15, focused (i.e., non-overlapping) retrieval mode—an average of merely 89 ms per CAS query and 49 ms per CO query over the 4.39 GB INEX Wikipedia collection.

Future Directions

With the already existing support for thesaurus/ontology-based query expansion (either in the form of classic query expansion, or using our incremental merging techniques for XML [13, 15]), we have a complete support for the W3C XPath 2.0 Full-Text specification, including phrase and proximity-aware keyword search (see also Subsection 31.9.2) and possibly non-monotonic score aggregations [4, 9]. Thus, our long-term goal for TopX will be to move beyond XPath and towards a full-fledged top-k engine for XQuery Full-Text. Further theses could start exploring XQuery support for TopX, which should serve as an initial step towards a true, ranking-aware, top-k algebra for XQuery Full-Text. The step from XPath-FT to XQuery-FT will require a notion of query plans and operator trees in TopX. As XQuery

[4] http://drops.dagstuhl.de/portals/index.php?semnr=08111
[5] http://www.inex.otago.ac.nz/

allows for the specification of multiple XPath expressions per XQuery expression, this poses new challenges for a holistic (i.e., for the entire XQuery expression), top-k-aware query optimizer.

References

- [1] M. AbuJarour. Efficient XML query processing and full-text search. Masters thesis, Universität des Saarlandes, 2008.

[2] H. Bast, D. Majumdar, R. Schenkel, M. Theobald, and G. Weikum. IO-Top-k at TREC 2006: Terabyte Track. In E. M. Voorhees and L. P. Buckland, eds., *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*, Gaithersburg, Maryland, 2006, pp. 551–555. NIST.

[3] H. Bast, D. Majumdar, R. Schenkel, M. Theobald, and G. Weikum. IO-Top-k: Index-access optimized top-k query processing. In U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim, eds., *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB 2006)*, Seoul, Korea, 2006, pp. 475–486. ACM. Acceptance ratio 1:7.

- [4] A. Broschart, R. Schenkel, and M. Theobald. Proximity-aware scoring for XML retrieval. In S. Geva, J. Kamps, and A. Trotman, eds., *Preproceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, Schloss Dagstuhl, Germany, 2008, pp. 46–49.

- [5] A. Broschart, R. Schenkel, M. Theobald, and G. Weikum. TopX @ INEX 2007. In N. Fuhr, M. Lalmas, and A. Trotman, eds., *Preproceedings of the 6th International Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2007)*, Schloss Dagstuhl, Germany, 2007, pp. 84–93. Springer.

- [6] C. Gherbaoui. Similarity measures for query expansion in TopX. Diploma thesis, Universität des Saarlandes, 2008.

- [7] L. Kasradze. Implementation of a file-based indexing framework for the TopX search engine. Masters thesis, Universität des Saarlandes, 2008.

- [8] O. Sammodi. Incremental relevance feedback for TopX. Masters thesis, Universität des Saarlandes, 2008.

- [9] R. Schenkel, A. Broschart, S. Hwang, M. Theobald, and G. Weikum. Efficient text proximity search. In N. Ziviani and R. A. Baeza-Yates, eds., *14th String Processing and Information Retrieval Symposium (SPIRE 2007)*, Santiago, Chile, 2007, *LNCS 4726*, pp. 287–299. Springer.

- [10] M. Stadtmüller. An XPath 2.0 full-text query parser for the TopX search engine. Bachelor thesis, Universität des Saarlandes, 2007.

[11] M. Theobald. *Efficient Top-k Query Processing for Text, Semistructured, and Structured Data.* Phd thesis, Universität des Saarlandes, 2006.

- [12] M. Theobald, M. AbuJarour, and R. Schenkel. TopX 2.0 at the INEX 2008 Efficiency Track. In S. Geva, J. Kamps, and A. Trotman, eds., *Preproceedings of the 7th Int. Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, Schloss Dagstuhl, Germany, 2008, pp. 230–244.

- [13] M. Theobald, H. Bast, D. Majumdar, R. Schenkel, and G. Weikum. TopX: Efficient and versatile top-k query processing for semistructured data. *The VLDB Journal*, 17(2):81–115, 2008.

• [14] M. Theobald and R. Schenkel. Overview of the INEX 2008 Efficiency Track. In S. Geva, K. Kamps, and A. Trotman, eds., *Preproceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, Schloss Dagstuhl, Germany, 2008, pp. 208–219.

[15] M. Theobald, R. Schenkel, and G. Weikum. Efficient and self-tuning incremental query expansion for top-k query processing. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, eds., *28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, Salvador, Brazil, 2005, pp. 242–249. ACM. Acceptance ratio 1:5.

[16] M. Theobald, R. Schenkel, and G. Weikum. An efficient and versatile query engine for TopX search. In K. Böhm, C. S. Jensen, L. M. Haas, M. L. Kersten, P.-A. Larson, and B. C. Ooi, eds., *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005)*, Trondheim, Norway, 2005, pp. 625–636. ACM. Acceptance ratio 1:6.

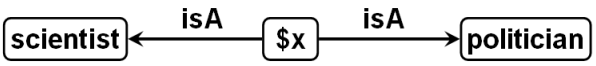
• [17] M. Theobald, R. Schenkel, and G. Weikum. TopX - efficient and versatile top-k query processing for text, semistructured, and structured data. In A. Kemper, H. Schöning, T. Rose, M. Jarke, T. Seidl, C. Quix, and C. Brochhaus, eds., *12. GI-Fachtagung Datenbanksysteme in Business, Technologie und Web (BTW 2007)*, Aachen, Germany, 2007, *Lecture Notes in Informatics*, vol. 103, pp. 475–485. Gesellschaft für Informatik.

• [18] M. Theobald, R. Schenkel, and G. Weikum. The TopX DB&IR engine (demo). In N. Koudas, ed., *2007 ACM SIGMOD International Conference on Management of Data*, Beijing, 2007, pp. 1141–1143. ACM.

31.6.2 NAGA: Graph Search with Statistics-based Ranking

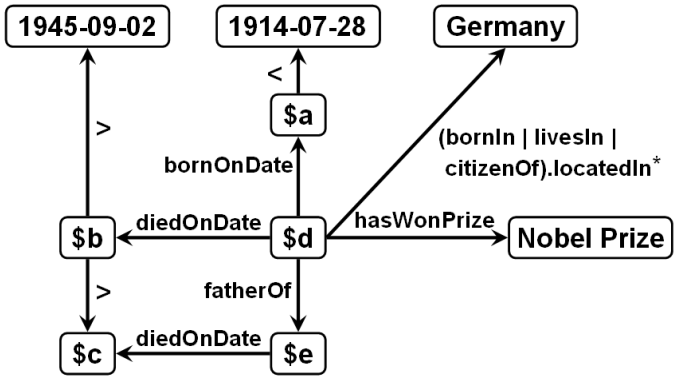
Investigators: Gjergji Kasneci, Shady Elbassuoni, Maya Ramanath, Fabian Suchanek, and Gerhard Weikum

Current keyword-oriented search engines for the World Wide Web do not allow for specifying *semantic* aspects of information needs. As an example, suppose we want to find a list of scientists who are also politicians. First, it is close to impossible to formulate this query in terms of keywords. Second, an answer to this question is probably distributed across multiple pages, so that no state-of-the-art search engine will be able to find it. In a more convenient setting, the user would have the possibility to specify entities and concepts corresponding to query terms as well as relations holding between them. The underlying engine would understand the query semantics, and would return ranked answers. In such a setting, the preceding query could be expressed in the following graphical form:



We address the above problem with NAGA [3, 4], a new semantic search engine, which operates on large knowledge graphs. The nodes of such graphs represent entities, i.e. persons, locations, movies, etc.; their edges represent binary relationship instances (or facts). Examples of facts could be *Max_Planck* BORNONDATE *1858-04-23* or *Max_Planck* TYPE *physicist*. These facts can be stored in a database as RDF triples, or relational tuples, or XML fragments with XLinks, etc. In order to query such a knowledge graph, NAGA provides a flexible query model based on entities and relationships as well as an effective ranking mechanism.

A NAGA query is a graph the nodes of which are labeled either with entity names or with variables, and the edges of which are labeled with regular expressions over relationship names. The regular expressions over relationship names allow NAGA's query language to go beyond SPARQL, the standard query language for Web ontologies stored in RDF/RDFS. A rather complex NAGA query asking for a *Nobel Prize* winner from *Germany* who survived both world wars outlived one of his children is depicted below.



Given a query graph q, NAGA tries to find all subgraphs in the knowledge graph that match q. These subgraphs are ranked based on confidence and informativeness. While the confidence intuitively captures the accuracy of results, the informativeness captures their prominence. To capture these criteria, we apply the principles of statistical language models (e.g., [6, 2, 8]), which are widely used for document retrieval, to the new setting of knowledge graphs. In general, the score of an answer graph $g = f_1, f_2, ..., f_k$ given a query graph $q = q_1, q_2, ..., q_k$ is computed as

$$Sc(g,q) = \prod_{i=1}^{k}(\lambda P(q_i|f_i) + (1-\lambda)P(q_i)) \approx \prod_{i=1}^{k}\frac{P(q_i|f_i)}{P(q_i)}$$

where f_i represents a single fact in the answer graph and q_i represents the query subgraph matched by f_i. The above score measure can be approximated by $KL(LM(q)|LM(g))$, i.e., the Kullback–Leibler divergence between the language model of q and the language model of g. In practice, this resolves to estimating occurrence statistics for each fact of an answer graph. Answers with higher fact occurrences are considered as more informative. The occurrence statistics for the facts can be estimated on the corpus from which they were extracted. We estimate them for each fact by the number of pages in the corpus in which we encounter that fact. Finally, these occurrence statistics can be weighted with additional confidence values for facts (if provided by the ontology).

NAGA's superior result quality was shown in an extensive user evaluation. For a subset of queries from the question answering benchmarks of TREC 2005 and TREC 2006, we compared NAGA's results with those returned by Google, Yahoo! Answers, and the question answering system START (`http://start.csail.mit.edu/`). NAGA excelled on the largest part of queries by fully leveraging its expressive query language and its powerful ranking mechanism. NAGA was recently extended by a new feature which allows users to discover closest relations between multiple entities using Steiner trees (see Subsection 31.6.3).

References

[1] J. Graupmann, R. Schenkel, and G. Weikum. The SphereSearch engine for unified ranked retrieval of heterogeneous XML and Web documents. In K. Böhm, C. S. Jensen, L. M. Haas, M. L. Kersten, P.-Å. Larson, and B. C. Ooi, eds., *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005)*, Trondheim, Norway, 2005, pp. 529–540. ACM. Acceptance ratio 1:6.

[2] D. Hiemstra. A probabilistic justification for using tf x idf term weighting in information retrieval. *Int. J. on Digital Libraries*, 3(2):131–139, 2000.

- [3] G. Kasneci, F. Suchanek, G. Ifrim, S. Elbassuoni, M. Ramanath, and G. Weikum. NAGA: Harvesting, searching and ranking knowledge (demo). In J. Tsong-Li Wang, ed., *Proceedings of the ACM SIGMOD 2008 International Conference on Management of Data (SIGMOD 2008)*, Vancouver, Canada 2008, 2008, pp. 1285–1288. ACM.

- [4] G. Kasneci, F. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. In *24th International Conference on Data Engineering (ICDE 2008)*, Cancun, Mexico, 2008, pp. 953–962. IEEE Computer Society.

- [5] G. Kasneci, F. Suchanek, M. Ramanath, and G. Weikum. How NAGA uncoils: Searching with entities and relations. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, New York, NY, USA, 2007, pp. 1167–1168. ACM Press.

[6] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1998, pp. 275–281. ACM.

- [7] F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 697–706. ACM.

[8] C. Zhai. Statistical language models for information retrieval: A critical review. *Foundations and Trends in Information Retrieval*, 2(3):137–213, 2008.

31.6.3 STAR: Top-k Steiner Trees for Graph-Search Ranking

Investigators: Gjergji Kasneci, Maya Ramanath, Mauro Sozio, and Fabian Suchanek

Large graphs and networks are abundant in modern information systems: entity-relationship graphs over relational data or Web-extracted entities, biological networks, social online communities, knowledge bases, and many more. Often such data comes with expressive node and edge labels that allow an interpretation as a semantic graph, with edge weights representing the strength of a semantic relation between two such entities. Discovering close relationships between two, three, or more entities (which are represented by graph nodes) is an important building block for many search, ranking, and analysis tasks. Real-life scenarios could include finding the closest relation between k given proteins (e.g., for medical reasons), the relation between k criminals, the most relevant data shared by k users, etc.

From an algorithmic point of view, the above task translates to computing the best Steiner trees between the given nodes, a classical NP-hard problem, that has been widely addressed in the literature [1, 2, 3, 4, 7]. In this setting, the total cost of a solution (i.e., a tree) is given by the sum of its edge weights. We address the problem with a new approximation

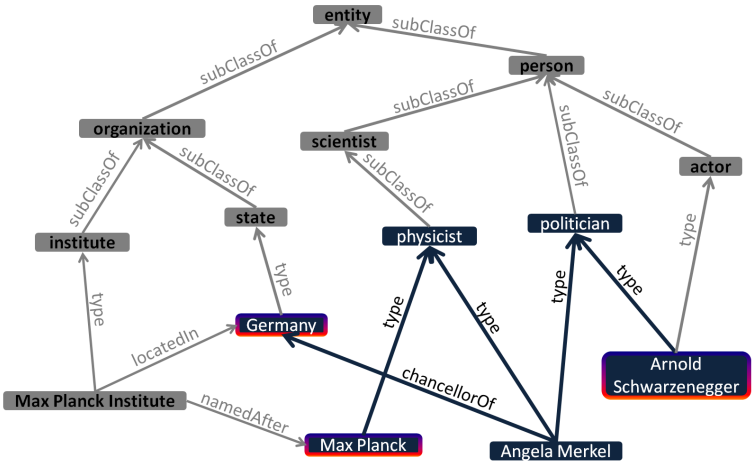


Figure 31.9: Steiner tree embedded into an example entity-relationship graph

algorithm, coined STAR [5], for relationship queries over large relationship graphs. STAR exploits the fact that often such graphs come with a taxonomic backbone: the *IsA*-hierarchy on the entities. Figure 31.9 depicts a sample entity-relationship subgraph. The bold edges represent a Steiner tree that interconnects the query entities *Max Planck*, *Germany*, and *Arnold Schwarzenegger*.

STAR processes a query by first building an initial tree that contains all query entities based on this hierarchy. In the above sample graph, such a tree can be found quickly, by following only the edges labeled with *type* and *subClassOf*. The taxonomic tree rooted at the *entity* node would be the initial tree for the query entities *Max Planck*, *Germany*, and *Arnold Schwarzenegger*. However, such a tree can be relatively large, and hence quite poor in terms of the underlying cost function. Consequently, in a second phase, STAR aims to iteratively improve the current solution by searching in its local neighborhood for better solutions. This works by removing certain paths that connect two subtrees in the current solution and replacing them by shorter ones from the underlying graph. Algorithm 1 gives a high-level overview of STAR's second phase. The set $LP(T)$ represents a priority queue containing all candidate paths that can be replaced in the current solution T. Heuristically, STAR aims to replace the longest of these paths by a shorter one from the underlying graph. When such a shorter path is found, a new result T' is constructed.

The method $Replace(lp, T)$ (Line 4), exploits sophisticated shortest path heuristics to determine the shortest paths that connect the two subtrees which result from the removal of the path lp. There are two main heuristics which guide the search process for the shortest path. The first heuristic aims to reduce the explored search space, by prioritizing low-degree nodes and by balancing the number of visited nodes across the single-source-shortest-path iterators. The second one aims at reducing the cost for managing the data structures needed for the exploration of the search space. In fact, in each improvement step STAR uses only two single-source-shortest-path iterators to explore the search space. This is a considerable advantage over the BANKS-I algorithm [1], which uses a single-source-shortest-path iterator

Algorithm 1 *improveTree*(T, V')

```
1: priorityQueue Q = LP(T) //ordered by decreasing weight
2: while Q.notEmpty() do
3:    lp = Q.dequeue()
4:    T' ← Replace(lp, T)
5:    if w(T') < w(T) then
6:       T = T'
7:       Q = LP(T) //ordered by decreasing weight
8:    end if
9: end while
10: return T
```

for each query node. The improvement procedure stops when no better solution can be found.

Although the final result of the second phase is only locally optimal, we prove that it is an $O(log(n))$-approximation of the optimal Steiner tree, where n is the number of the query entities. This theoretical result has a very important implication. It entails that the approximation ratio for the cost of the final tree returned by STAR is independent of the tree constructed in the first phase. Furthermore, we show in our experiments that in practical cases, the results returned by STAR are qualitatively comparable to or even better than the results returned by a classical 2-approximation algorithm [7].

STAR can be generalized into a top-k Steiner tree algorithm in an almost natural way. This is done by relaxing the edge weights of the best result returned by STAR and by forcing the algorithm to explore new paths. This way, new trees of higher weight can be found. Furthermore, all trees produced through the iterative improvement strategy of STAR serve in this phase as possible top-k candidates. STAR has a pseudo-polynomial worst-case runtime. In practice, it produces results very efficiently. Our experiments show that in terms of efficiency STAR outperforms the best state-of-the-art database methods [1, 2, 3, 4] by a large margin. STAR is implemented as a query answering component of the NAGA [6] system (see Subsection 31.6.2).

References

[1] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In *ICDE*, 2002, pp. 431–440. IEEE Computer Society.

[2] B. Ding, J. X. Yu, S. Wang, L. Qin, X. Zhang, and X. Lin. Finding top-k min-cost connected trees in databases. In *ICDE*, 2007, pp. 836–845. IEEE.

[3] H. He, H. Wang, J. Yang, and P. S. Yu. BLINKS: ranked keyword searches on graphs. In C. Y. Chan, B. C. Ooi, and A. Zhou, eds., *SIGMOD Conference*, 2007, pp. 305–316. ACM.

[4] V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, and H. Karambelkar. Bidirectional expansion for keyword search on graph databases. In K. Böhm, C. S. Jensen, L. M. Haas, M. L. Kersten, P.-Å. Larson, and B. C. Ooi, eds., *VLDB*, 2005, pp. 505–516. ACM.

- [5] G. Kasneci, M. Ramanath, M. Sozio, F. Suchanek, and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *Proceedings of the 25th International Conference on Data Engineering (ICDE 2009)*, Shanghai, China, 2009. IEEE Computer Society.

- [6] G. Kasneci, F. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. In *24th International Conference on Data Engineering (ICDE 2008)*, Cancun, Mexico, 2008, pp. 953–962. IEEE Computer Society.

[7] L. T. Kou, G. Markowsky, and L. Berman. A fast algorithm for Steiner trees. *Acta Inf.*, 15:141–145, 1981.

31.6.4 Personalized Ranking

Investigators: Julia Luxenburger and Shady Elbassuoni

Search personalization has been pursued in many ways, in order to provide better overall search experience to individual users. However, blindly applying personalization to all user queries is not always appropriate. User interests change over time, a user sometimes works on very different categories of tasks within a short time span, and history based personalization may impede a user's desire to discover new topics. This work aims at answering the questions: (1) how do we decide whether a query is expected to benefit from prior history information and (2) how to estimate the user's search goal and selectively choose the best matching part of the history (whether immediate or long-term) in order to carry out personalization.

There are numerous attempts towards personalizing Web search. Most approaches make use of implicit user relevance feedback. They achieve personalization by either *re-ranking* of Web search results [8, 9] or *query expansion* [1]. However, both types of approaches are highly query-dependent. They might improve some queries, while harming others. Our earlier work [3, 4] aimed at more adaptive techniques by distinguishing the various user search modes. While our experimental results showed improvements for the search modes of re-finding known information and satisfying ad-hoc information needs, our personalization approach showed query-dependent variations and less impressive performance gains when the user explored topics of her general interest.

We thus propose a personalization framework that is selective in a two-fold sense. First, it selectively employs personalization techniques for queries that are expected to benefit from prior history information, while refraining from undue actions otherwise. Second, our framework is based on fine-grained statistical language models for units of a user's past search-and-browse behavior, coined *tasks*. They reflect possibly non-homogeneous, varying aspects of user interests and are obtained by means of a hierarchical clustering of the user's profile. The atomic units to be clustered are past user sessions which consist of cohesive *query chains*, i.e., subsequently posed queries including their result clicks and further browsed documents within the same session, as well as documents browsed by the user independently of a query. Thus, we represent the user profile as the full dendogram of tasks which reaches from small tasks consisting only of a single session, to the largest task encompassing the whole user search and browse history.

Selective Personalization Strategy. We distinguish two cases: either (1) the current query is the first query in the current session, or (2) there exists some query history already, and the current query is a refinement of a previously issued query in the same search session. In the first case, we retrieve the *top-k* tasks $T_1, \ldots, T_k$ which are most similar to the query from the user's profile. In the latter case, the tasks present in the user profile are accompanied by

a current task made up by all the actions of the currently active session, represented by the language model T_{k+1}.

Similarly, we perform a hierarchical clustering of the query's *result set* items (each represented by its title and snippet information) to obtain candidate query *facets* $F_1, \ldots, F_m$ which represent the different aspects the query might span. Each obtained query facet and each task is represented by a unigram language model.

We then consider the Kullback-Leibler (KL) divergence between a query facet F_i and a task T_j

$$KL(F_i||T_j) = \sum_w P(w|F_i).\log \frac{P(w|F_i)}{P(w|T_j)}$$

to determine the facet-task pair (F_i^*, T_j^*) with the lowest KL divergence. This way we learn the query facet F_i^* that is closest to the user's interests, and which represents the most probable meaning of the query in case of ambiguity. At the same time, the task T_j^* represents the best-matching part of the user profile to the query facet F_i^*. The KL divergence between F_i^* and T_j^* further characterizes the strength of their similarity. If $KL(F_i^*, T_j^*)$ is larger than a threshold σ, we conclude that the current query goes for a previously unexplored task, and thus refrain from biasing the search results. Otherwise, we either reformulate the query or re-rank the original search results as described in the following.

Means of Personalization. Here, we tackle the question of how to personalize once personalization has been deemed useful for the current query. Given a facet of the current query F_i^* and a task T_j^* which are most similar to each other, we update the query representation with those terms that are the best discriminators of the chosen query facet F_i^* from all other query facets, while being most similar to the task T_j^*. That is, we choose terms whose KL divergence between the union of the chosen facet-task pair and the remaining query facets is above a second threshold τ:

$$v(w) = P(w|F_i^* \cup T_j^*)\ \log \frac{P(w|F_i^* \cup T_j^*)}{P(w|\bigcup_{i \neq i^*} F_i)} < \tau$$

with

$$P(w|\bigcup_{i \neq i^*} F_i) = \frac{1}{|\{i|i \neq i^*\}|}\ \sum_{i \neq i^*} P(w|F_i)$$

Based on the KL divergence between the new query representation and the result items' language models estimated from their title and snippet information, we re-rank the original results. Thus, in case of ambiguity, the query is biased towards the facet with the best-matching counterpart in the user's profile.

Finally, we introduce a third threshold δ to allow for an automatic reformulation of the query whenever we find strongly discriminative terms. Terms for which $v(w) < \delta$ holds, qualify for query expansion, while terms with $\delta \leq v(w) < \tau$ and $P(w|\bigcup_i F_i) > 0$ holds, qualify for re-ranking the original top-50 search results. This way, the use of the more invasive query expansion is restricted to queries for which there is a certain level of confidence in its success.

Evaluation. To evaluate the effectiveness of our proposed approach we conducted an initial user study where we asked seven volunteers to install our personalization agent on their local machines, logging their search and browsing activities for a period of two months. After this phase of logging, each participant evaluated a set of self-chosen queries. In summary, our experimental results proved the benefits of our personalization framework by contrasting personalized results with the original search results. Especially, our selective and task-aware personalization scheme was shown to be superior over a simple bag-of-words personalization scheme [7]. By testing the different components of our personalization scheme separately, we were also able to show the individual advantages of each component [6]. Finally, our cross-validation results showed that the learned parameters of our framework generalized well [5].

References

[1] P.-A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the Web. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, eds., *SIGIR*, 2007, pp. 7–14. ACM.

- [2] M. Dudev, S. Elbassuoni, J. Luxenburger, M. Ramanath, and G. Weikum. Personalizing the search for knowledge. In *2nd International Workshop on Personalized Access, Profile Management, and Context Awareness: Databases (PersDB 2008)*, Auckland, New Zealand, 2008, pp. 1–8.
- [3] S. Elbassuoni. Adaptive personalization of Web search. Masters thesis, Universität des Saarlandes, 2007.
- [4] S. Elbassuoni, J. Luxenburger, and G. Weikum. Adaptive personalization of Web search. In K. Rodden, I. Ruthven, and R. W. White, eds., *Proceedings of the 1st Workshop on Web Information Seeking and Interaction*, Amsterdam, The Netherlands, 2007, pp. 1–5. ACM.
- [5] J. Luxenburger. *Modeling and Exploiting User Search Behavior for Information Retrieval*. Phd thesis, Universität des Saarlandes, 2008.
- [6] J. Luxenburger, S. Elbassuoni, and G. Weikum. Matching task profiles and user needs in personalized Web search. In *17th Conference on Information and Knowledge Management (CIKM 2008)*, Napa Valley, USA, 2008, pp. 689–698. ACM.
- [7] J. Luxenburger, S. Elbassuoni, and G. Weikum. Task-aware search personalization. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, eds., *31st Annual International ACM SIGIR Conference (SIGIR 2008)*, Singapore, 2008, pp. 721–722. ACM.

[8] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, and W. Teiken, eds., *CIKM*, 2005, pp. 824–831. ACM.

[9] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, eds., *SIGIR*, 2005, pp. 449–456. ACM.

31.6.5 Ranking for Temporal Keyword Queries

Investigators: Irem Arikan, Srikanta Bedathur, Klaus Berberich, and Gaurav Pandey

Existing relevance models used in Information Retrieval, such as Okapi BM25, often fail to produce satisfactory results for information needs that have a strong temporal component. Consider, as an example, a journalist who wants to find out about activities of the European

Commission in the 1990s and thus issues the keyword query "european commission activities 1990s" to a search engine. In existing relevance models, a document is considered highly relevant, if it contains many occurrences of all –or at least most– query keywords. However, a document that contains detailed information about the European Commission's activities on specific dates in the 1990s (e.g., the resigning of the Santer Commission on March 15, 1999) but that does not contain the term "1990s" would not be deemed highly relevant and therefore would not be shown to our journalist.

This is because the semantics inherent to temporal expressions, such as *on March 15, 1999* or *between 1992 and 1997*, remain hidden to the relevance model as noted in [1]. Even though both temporal expressions refer to times that lie within our period of interest, they would not be matched to the query keyword "1990s" by existing approaches.

In this line of work we focus on overcoming this problem by making relevance models aware of temporal expressions and their inherent semantics. Our objective is to improve retrieval effectiveness, and thus user satisfaction, for temporal information needs.

Our approach extends existing relevance models, more precisely the class of statistical language models, such as the seminal approach by Ponte and Croft [4]. For each document in the collection, these approaches determine a generative model based on the terms contained in the document and their respective frequencies. When given a user query q, the probability $P(q|d)$ that the query is produced from the generative model associated with document d is estimated and used as an indicator of the document's relevance to the query.

We advance beyond this state-of-the art by making temporal expressions first-class citizens of the relevance model. In our model, we distinguish between the textual part d_{tx} and the temporal part d_{te} of documents. The former is comprised of regular terms (e.g., "european", "union", and "commission" in our above example), whereas the latter is comprised of extracted temporal expressions (e.g., *on March 15, 1999*). Analogously, when given a user query, we extract temporal expressions and yield a textual query part q_{tx} and a temporal query part q_{te}. The relevance of a document to the user query is then estimated as

$$P(q|d) = P(q_{tx}|d_{tx}) \times P(q_{te}|d_{te}) \ .$$

In the above formula, the first part $P(q_{tx}|d_{tx})$ is the probability that the textual part of the query is generated from the textual part of the document, which we determine using the aforementioned Ponte and Croft [4] approach. The second part $P(q_{te}|d_{te})$, on the other hand, is the probability that the temporal expressions contained in the user query are generated from the temporal part of the document, which we estimate using our novel generative model for temporal expressions described in the following.

Our approach presented in [3] maps temporal expressions to time intervals on the time axis. The lower part of Figure 31.10 shows a concrete example of how the temporal expression *in March, 1999* is mapped to its corresponding time interval. Our generative model for temporal expressions consists of two steps. In a first step, a temporal expression is drawn uniformly at random from the temporal expressions contained in the document. Following that, in a second step, we generate a temporal expression from the temporal expression drawn in the first step. Here, we utilize our time-interval representation, i.e., we generate another time interval $[b', e']$ from the time interval $[b, e]$ corresponding to the temporal expression drawn. As a concrete example, the upper part of Figure 31.10 shows four time intervals with their

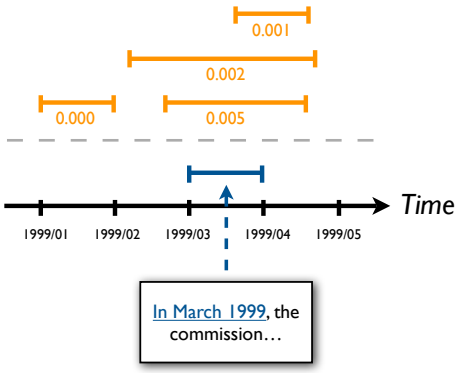


Figure 31.10: Generative model for temporal expressions

respective probabilities of being generated from the temporal expression *in March, 1999*. The generation of $[b', e']$ is done such that (i) $[b', e'] \cap [b, e] \neq \emptyset$ (i.e., we assure that an overlapping time interval is generated), (ii) time intervals whose boundaries are closer to the boundaries of $[b, e]$ have a higher likelihood of being generated (i.e., we favor time intervals that deviate less from $[b, e]$).

Putting these building blocks together, we obtain a relevance model aware of temporal expressions and their inherent semantics. We investigated the effectiveness of our approach using the English Wikipedia as a real-world dataset. A user study, whose details are given in [2], showed that our approach yields a significant improvement in retrieval effectiveness for temporal information needs over a text-only language-model approach.

Our recent paper [3] was awarded the Best Late-Breaking Result in the Second ACM International Conference on Web Search and Data Mining (WSDM '09).

References

[1] O. Alonso, M. Gertz, and R. Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35–41, 2007.

- [2] I. Arikan. Exploiting temporal references in text retrieval. Masters thesis, Universität des Saarlandes, 2009.
- [3] I. Arikan, S. Bedathur, and K. Berberich. Time Will Tell: Leveraging temporal expressions in IR. In *Second ACM International Conference on Web Search and Data Mining (WSDM '09) - Late Breaking Results*, Barcelona, Spain, 2009. ACM.

[4] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1998, pp. 275–281. ACM.

31.6.6 U-RDF: Efficient Reasoning in Uncertain RDF Knowledge Bases

Investigators: Martin Theobald, Mauro Sozio, and Fabian Suchanek

Ontological knowledge representations, in particular automatically extracted open-domain Web ontologies like DBpedia [1] (`http://dbpedia.org/`), KnowItAll [5] (`http://knowitall.`

org/), or Intelligence-In-Wikipedia [10] inherently exhibit a high degree of uncertainty. This arises from various forms of semi-automatic information extraction, data integration and fusion of different ontological sources, as well as pattern mining and natural language processing. Depending on the knowledge extraction tools used, but also due to false information delivered by different Web sources, many contradicting facts may be derived and yet need to be captured concisely by the underlying representation formalism. Often these inconsistencies are not obvious at first sight but could only be uncovered through intricate and possibly expensive inference steps. These intricate relationships among data objects, as well as the lack of a common schema, call for the use of graph-based, schema-less representation models like RDF/RDFS, as well as novel query processing techniques for SPARQL, the standard query language for RDF.

One way of dealing with inconsistencies is to prune as much noise as possible upfront and to detect possible inconsistencies already during the extraction process. However, this would be at the cost of a possible loss not only of noisy data but also of a great amount of potentially valuable data. Another, maybe even more intriguing, option is to gather as much data as possible and to leave the task of managing uncertainty and inconsistencies up to the knowledge base. Pursuing the first strategy results in a static and largely consistent knowledge base as it is our goal with the YAGO/SOFIE setting described earlier (see Section 31.4). Following the latter strategy, on the other hand, opens a variety of challenges to *dynamically* manage inconsistencies and perform uncertain reasoning directly at query processing time, in order to present a concise snapshot of the possible instances of the real world in direct response to a user's query.

In such open-domain knowledge representations, classical evaluation strategies based on "hard" propositional logic and Boolean evaluation would most of the times get lost in expensive satisfiability tests just to finally return "unsatisfiable" and yield empty results, as uncertain knowledge extracted from the Web is inherently not suitable for these strict verification methods. "Softer" evaluation and reasoning on uncertainty is key for the flexible evaluation of structured queries over graph-based, uncertain data. At the same time, this naturally conveys a well-defined result ranking to easily identify the most likely results in a probabilistic interpretation of the input confidences. Valid query results no longer have to be exact matches to all the constraints involved in both the knowledge representation and the inferencing rules, rather we prefer those results that satisfy most of the constraints with a high confidence, possibly by violating only a few of the low-confidence constraints.

A Data Model and Query Processing Framework for Uncertain RDF Data

With U-RDF, we are currently working on a novel data model for representing uncertainty in non-schematic, graph-based RDF knowledge bases, based on a formal *possible worlds* semantics (see, e.g., [2, 3, 4, 8]), and a probabilistic extension to SPARQL, the W3C query language for RDF/RDFS. To the best of our knowledge, our data model is the first to combine representation models from classic probabilistic databases, thus capturing *tuple-level uncertainty* and *hard key constraints* [4], with an efficient first-order reasoning framework over *soft rules* in the form of Prolog/Datalog-like Horn clauses [6, 7, 9].

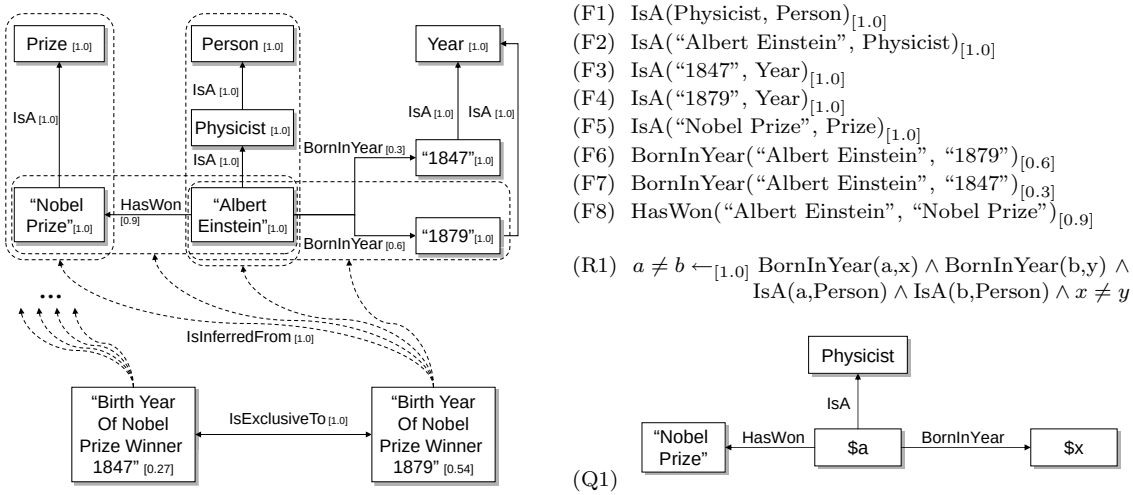


Figure 31.11: Example with uncertain facts $(F1)$-$(F8)$, inferencing rule $(R1)$, and query $(Q1)$

Example. *Consider a small knowledge base captured by facts* $(F1)$-$(F8)$ *and inference rule* $(R1)$*, as well as a query* $(Q1)$*, which is intended to return birth years of physicists who won the Nobel prize, as depicted in Figure 31.11. On the left, this data appears in graph-like form, and the facts underlying the graph appear on the right. Observe that each fact has a level of confidence, e.g.,* $(F1)$-$(F5)$ *have a perfect confidence of 1.0 (perhaps they were handcrafted), whereas* $(F6)$-$(F8)$ *have a lower degree of confidence (perhaps they were automatically derived). Moreover,* $(R1)$ *states that a person cannot have two different dates-of-birth. The confidence weight associated with the above implication indicates that the rule itself may be uncertain, which should also be taken into account when computing the confidences of results derived by applying* $(R1)$*.*

Note that $(R1)$*, together with* $(F1)$-$(F4)$*, and by transitively exploiting the* `IsA` *relation,* implicitly *imposes the mutual exclusiveness of the latter two facts* $(F6)$ *and* $(F7)$*. It is the role of the inference mechanism of the knowledge base to make this information* explicit*, by applying* $(R1)$ *to the subgraph that is relevant to* $(Q1)$*. That is, the evaluation of* $(Q1)$ *and a respective materialization of the two possible results (coined "reification" in RDF terminology) extends the RDF graph by two new entities, depicted as the lower two nodes in the Figure 31.11. These new nodes are connected via a special* `IsExlusiveTo` *edge that also forms the desired (and explicit) fact* $(F9)$ *IsExclusiveTo(BirthYearOfNobelPrizeWinner1847, BirthYearOfNobelPrizeWinner1879)*$_{[1.0]}$*. Furthermore, different* `IsInferredFrom` *edges can be used to represent the lineage information of the new nodes in the RDF graph. The* `IsInferredFrom` *edges may directly connect result entities and/or facts with other facts, thus yielding "facts over facts". Here, node weights are introduced to capture the result confidences of a previous query.* □

An Efficient Approximation Algorithm for Weighted MAX-SAT

The U-RDF data model exhibits many analogies to graphical probabilistic models like Markov Networks; in particular the combination with first-order inferencing rules makes it

very similar to recent related works on Markov Logic Networks [9]. In U-RDF, however, the reasoning can be triggered *selectively*, by restricting it to the specific subgraph in the knowledge base that is relevant to a query. In particular the Datalog-style inferencing rules are evaluated only over connected components of the knowledge base that are related to at least one fact that matches a query atom. Moreover, U-RDF considers efficient query processing issues over a combination of soft and hard rules, in contrast to existing approaches considering either only hard or only soft constraints [4, 7, 9] in isolation. Similarly to uncertainty and lineage databases (coined "ULDBs" [3]), the explicit storage of *data lineage* allows for a "closed and complete" representation model. Here, RDF yields a very convenient mechanism to materialize query results and lineage information directly as new facts into the graphical knowledge base, i.e., through reification of facts into a single RDF resource that may be used to represent also more complex (n-ary) relationships. Thus, U-RDF aims to overcome the limitations in the expressiveness of probabilistic database systems, where fixed schemata and strong independence assumptions among the data objects are still prevalent. Similarly to the mechanism we developed for consistency checks in SOFIE (see Subsection 31.4.2), the core of our approach is a linear-time approximation algorithm for a generalized version of the Weighted MAX-SAT problem with tight approximation guarantees. This Weighted MAX-SAT setting provides a natural way of dealing with inconsistencies between the underlying knowledge base and our inferencing rules.

Our long-term research goal for the U-RDF framework will be to combine tractable subproblems in probabilistic databases and logics-based reasoning, and to map these onto scalable and mature database technology. Various forms of "soft" (both rule-based and statistical) and "hard" inference techniques are intended as a new core functionality of the database system. Even in the non-probabilistic setting, query processing often needs to traverse large parts of the graph, when query predicates are not selective, while confidence computations over probabilistic data are known to be #P-complete [4]. Thus, efficient, top-k-style, early-termination algorithms already applied to probabilistic databases [8] will be of high interest in this novel domain as well.

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. F. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, eds., *ISWC/ASWC*, 2007, *LNCS 4825*, pp. 722–735. Springer.

[2] D. Barbará, H. Garcia-Molina, and D. Porter. The management of probabilistic data. *IEEE Trans. Knowl. Data Eng.*, 4(5):487–502, 1992.

[3] O. Benjelloun, A. D. Sarma, A. Y. Halevy, M. Theobald, and J. Widom. Databases with uncertainty and lineage. *VLDB J.*, 17(2):243–264, 2008.

[4] N. N. Dalvi and D. Suciu. The dichotomy of conjunctive queries on probabilistic structures. In L. Libkin, ed., *PODS*, 2007, pp. 293–302. ACM.

[5] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the Web. *Commun. ACM*, 51(12):68–74, 2008.

[6] N. Fuhr. Probabilistic Datalog - a logic for powerful retrieval methods. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995, pp. 282–290. ACM.

[7] A. Jha, V. Rastogi, and D. Suciu. Query evaluation with soft-key constraints. In M. Lenzerini and D. Lembo, eds., *PODS*, 2008, pp. 119–128. ACM.

[8] C. Re, N. N. Dalvi, and D. Suciu. Efficient top-k query evaluation on probabilistic data. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE)*, 2007, pp. 886–895.

[9] M. Richardson and P. Domingos. Markov Logic Networks. *Machine Learning*, 62(1-2):107–136, 2006.

[10] D. S. Weld, F. Wu, E. Adar, S. Amershi, J. Fogarty, R. Hoffmann, K. Patel, and M. Skinner. Intelligence in Wikipedia. In D. Fox and C. P. Gomes, eds., *AAAI*, 2008, pp. 1609–1614. AAAI Press.

31.7 Query Processing and Optimization

Coordinator: Thomas Neumann

Efficient execution of queries on large data collections is a critical issue for most database systems. Even though hardware performance increases, data grows accordingly, and maintaining good query response times and high throughput is a challenging problem relevant for wide classes of applications. Query processing itself is concerned with providing efficient primitives for formulating and executing queries inside a database system, while query optimization tries to find the most efficient combination of execution primitives for a given user query. Both parts interact with each other to provide efficient query execution, exploiting the given application characteristics.

31.7.1 RDF-3X: Scalable RDF Querying

Investigators: Thomas Neumann and Gerhard Weikum

RDF (Resource Description Framework) is a data representation format for schema-free structured information that is increasingly gaining momentum in the context of Semantic-Web corpora, life sciences, and also Web 2.0 platforms. While originally proposed for the Semantic Web, it has gained popularity for building large-scale data collections. In particular biologists and other life scientists like RDF because of its ease of use and flexibility. They can easily collect data without a schema-first database design phase – a paradigm known as "pay-as-you-go" dataspaces [2]. Moreover, it is easy to add metadata, annotations, and lineage information, and represent all of this uniformly together with the primary data.

RDF data collections can be seen as entity-relationship graphs with edges corresponding to *(subject, property, object) (SPO) triples* (where *property* is often also referred to as *predicate*). For example, the fact that the French novelist (and Nobel prize winner) Jean-Marie Le Clezio has written the book "Desert" would be represented by the following four triples:
`(Id1, hasName, Jean-Marie Le Clezio), (Id1, hasNationality, French),`
`(Id1, hasWritten, Id2), (Id2, hasTitle Desert).`
Querying such RDF data amounts to evaluating graph patterns, expressed in the SPARQL query language. For example, the query with the following three *triple patterns*

`?x hasName ?y . ?x hasNationality French . ?x hasProfession novelist`
finds the names of all French novelists (where the dot denotes a conjunction). During execution, the database system has to find all possible variable bindings such that the resulting triples occur in the database instance.

Executing these queries is very challenging due to the lack of schema information and the fine-grained nature of RDF. Everything is expressed as triple and when evaluating multiple triple patterns, each scanned triple could potentially join which each other triple. From a relational point-of-view, an RDF database consists of one big relation of triples, and queries are formulated as a large number of self-joins with additional selection predicates. Both query processing and query optimization is non-trivial for this scenario, and the standard techniques give relatively poor performance for more complex queries.

We developed the open source RDF-3X system [3] for efficient storage and retrieval of RDF data. It provides efficient RDF processing based upon four main components:

- First, it stores all triples in very space-efficient triple indexes using compressed B^+-trees. RDF triples are converted into integer triples using dictionary compression, and these integer triples are then stored in all $3! = 6$ possible permutations in B^+-trees with leaf compression. This storage technique allows RDF-3X to answer each individual triple pattern using a single range scan in a clustered B^+-tree, which is very fast. Additionally, this exhaustive indexing allows for retrieving data items in any sort order, which allows for efficient merge joins when combining results. Moreover, aggregated indexes based upon attribute projections allow for faster reading when only parts of a triple are including in the query result.

- Second, a RISC-style execution engine tuned towards very fast merge joins combines the triple patterns and eliminates tuples as quickly as possible, leaving the more expensive hash joins to later stages with reduced tuple counts.

- Third, RDF-3X maintains histograms based upon hierarchical triple aggregation and frequent paths occurring in the database instance to predict the number of qualifying triples for triple patterns and join results.

- Finally, a smart query optimizer chooses the cost-optimal join ordering for a given query using these statistics and builds an efficient execution plan exploiting all 15 available indexes.

In experiments on large data sets, RDF-3X outperformed previous approaches by an order magnitude, both due to more efficient query processing and smarter construction of execution plans.

These techniques work very well for data sets with millions of triples but are still not sufficient when data sets contain billions of triples. While data sets of this size are still uncommon for RDF data, some well known collections like the UniProt RDF protein database[6] reach this size. At this size, evaluating even a single triple pattern becomes prohibitively expensive, as individual triple patterns tend to be unselective and the I/O costs of reading large fractions of such huge data sets are very high. To solve this problem, we developed

[6] `http://dev.isb-sib.ch/projects/uniprot-rdf/`

a flexible and low-overhead, sideways-information-passing (SIP) technique for RDF-3X [5] that propagates domain information across the execution plan: both scan and join operators record the attribute bindings they observe, and use these to construct (a superset of) the domain for each attribute occurring in the query. Using this observed domain information, scan operators can decide if the currently examined triples have a chance to make it into the final result, and can skip large parts of the data if they deduce that they do not qualify. This SIP technique exploits the structure of the query execution plan (e.g., merge joins vs. hash joins, etc.) and uses the known access patterns of the operators to constantly refine (i.e., shrink) the potentially relevant domain for each attribute. Another problem in this context are selectivity estimations for very large data sets. Regular histograms perform poorly, and it is unclear how to scale them as the data grows. Instead, we used the existing aggregated indexes for selectivity estimation and build precomputed compressed indexes with join cardinalities for very accurate prediction of result cardinalities. Both techniques together improved the RDF-3X performance on very large data sets by more than an order of magnitude. According to the performance numbers in published experimental studies (e.g., [1]), RDF-3X is currently the fastest system for RDF data with full support for SPARQL-style triple patterns.

References

[1] D. J. Abadi, A. M. 0002, S. Madden, and K. J. Hollenbach. Scalable Semantic Web data management using vertical partitioning. In C. Koch, J. Gehrke, M. N. Garofalakis, D. Srivastava, K. Aberer, A. Deshpande, D. Florescu, C. Y. Chan, V. Ganti, C.-C. Kanne, W. Klas, and E. J. Neuhold, eds., *VLDB*, 2007, pp. 411–422. ACM.

[2] A. Y. Halevy, M. J. Franklin, and D. Maier. Principles of dataspace systems. In *PODS*, 2006, pp. 1–9.

- [3] T. Neumann and G. Weikum. RDF-3X: a RISC-style engine for RDF. *Proceedings of the VLDB Endowment*, 1(1):647–659, 2008.
- [4] T. Neumann and G. Weikum. The RDF-3X engine for scalable management of RDF data. Research Report MPI-I-2009-5-003, Max-Planck-Institut für Informatik, Saarbrücken, Germany, 2009.
- [5] T. Neumann and G. Weikum. Scalable join processing on very large RDF graphs. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD 2009)*, Providence, USA, 2009. ACM.

31.7.2 Join Order Optimization

Investigator: Thomas Neumann in cooperation with Guido Moerkotte

Programs use declarative query languages to retrieve data from a database system. The query specifies what the application is interested in, but not how the the result should be computed. The database system is free to choose between different execution alternatives and it uses this degree of freedom to find the most efficient execution strategy for the given query.

One of the most important optimization problems in this context is *join order optimization*. Join operators are used very frequently, occurring in nearly all queries. They are both freely reorderable, as they are associative and commutative, and – individually – relatively expensive, which makes ordering joins both challenging and important. Changing the join order can effect query execution times by orders of magnitude. In general, the problem of finding the optimal

join order is NP hard, but we showed in previous work that the difficulty strongly depends on the structure of the query [1]. Here, the structure of a query is the query graph derived from the query by creating nodes for all relations in the query and connecting two nodes if they have a join predicate in common. We developed a graph-based dynamic programming (DP) strategy that reduces the problem of finding the optimal join order to the problem of enumerating all connected subgraphs in the query graph. The resulting DPccp algorithm [1] meets the known lower bound for all enumeration-based DP algorithms and it significantly outperforms previous algorithms.

While the DPccp algorithm solves the classical join ordering very efficiently, it has problems with some unusual kinds of queries: some queries cannot be described by classical query graphs alone due to unusual join predicates, and thus cannot be optimized by DPccp. This is unfortunate, as commercial database systems must be able to handle any kind of syntactically valid query, and database vendors are unwilling to maintain two different optimizers for different kinds of queries. The problem arises from query predicates like, for example, $R_1.a + R_2.b = R_3.c + R_4 + d$. These predicates use attributes from more than two relations and therefore they imply edges between entire sets of nodes ($\{R_1, R_2\}$ and $\{R_3, R_4\}$ in the example), which cannot be expressed in regular graphs.

We addressed this problem by generalizing our graph based DPccp algorithm to hyper-graphs, i.e., graphs where edges connect sets of nodes. The resulting DPhyp algorithm [2] uses the same high-level idea as the original algorithm: We know that a set of relations can be combined into a join tree if and only if they are connected in the query graph. Thus we can solve the join ordering problem by enumerating all connected subgraphs of the query graph in a suitable order and producing join trees for all subgraphs. The main difficulty here is that "connectedness" is more complex in hyper-graphs, and that, therefore, incremental subgraph construction may require building non-connected intermediate steps, which complicates the algorithm. By using careful bookkeeping and exploiting the enumeration order of our DP strategy, we were able to lift the graph-based DP hyper-graphs, and consequently could efficiently optimize join queries with arbitrary join predicates [3].

A great advantage of supporting hyper-graph-structured query graphs is that we can now express much more complex queries than before, not just more complex join predicates. Some queries, in particular nested queries, use non-inner joins (e.g., outer joins or anti-joins) during query translation. These join operators are not freely reorderable like the regular joins, but only allow limited reordering. We integrated these non-inner join operators in the join ordering process [3] by first deriving all reordering restrictions from the original query, and then encoding these restrictions as hyper-edges into the query graph. This allows us to use the efficient DPhyp algorithm for a very large class of problems, supporting the entire range of SQL join operators and predicates.

The DPhyp algorithm efficiently solves all kinds of join-ordering problems by adapting to the query graph structure, but ultimately the join-ordering problem remains NP hard. For typical queries with less than 10 relations this is not a problem, but some application scenarios use large (program-generated) queries with 20 or 50 relations. Here, DPhyp reaches its limits, as in general such large queries can no longer be optimized exactly due to runtime constraints. However, even here, the complexity depends on the query structure: If the query graph forms a chain of n relations (i.e., one relation is joined to the next one and so on) the runtime complexity of DPhyp is $O(n^3)$, while the runtime complexity is $O(n\ 2^n)$ if the query

graph forms a star (i.e., one relation is joined with all other relations). This means we can still optimize chain queries with 50 relations easily using DPhyp, while star queries become too expensive for much smaller n.

This observation led us to a new principle of optimizing very large queries: we first examine the query graph structure, and if the implied search space is too large, we keep *simplifying* the query by adding reordering restrictions in the form of hyper-edges [4] until we can solve the simplified problem exactly. The simplification is based on a heuristic, but the great advantage of this approach over classical heuristics for join ordering is that we do not have to construct join plans (which is hard to do using heuristics), but instead we forbid certain join plans. As we usually only have to add a few restrictions until the search space becomes tractable, we add these restrictions where we are most certain about their preservation of the optimum (i.e., the safe choices) first. Experiments have shown that this simplification strategy produces execution plans that outperform those of previous heuristical join ordering techniques by multiple orders of magnitude, and allows to handle all kinds of queries based upon graph-based DP.

References

[1] G. Moerkotte and T. Neumann. Analysis of two existing and one new dynamic programming algorithm for the generation of optimal bushy join trees without cross products. In U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim, eds., *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB 2006)*, Seoul, Korea, 2006, pp. 930–941. ACM. Acceptance ratio 1:7.

- [2] G. Moerkotte and T. Neumann. Dynamic programming strikes back. In J. T.-L. Wang, ed., *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2008)*, Vancouver, Canada, 2008, pp. 539–552. ACM.

- [3] G. Moerkotte and T. Neumann. Faster join enumeration for complex queries. In *Proceedings of the 24th International Conference on Data Engineering (ICDE 2008)*, Washington, USA, 2008, pp. 1430–1432. IEEE.

- [4] T. Neumann. Query simplification: Graceful degradation for join-order optimization. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD 2009)*, Providence, USA, 2009. ACM.

- [5] T. Neumann and G. Moerkotte. A framework for reasoning about share equivalence and its integration into a plan generator. In *Datenbanksysteme in Business, Technologie und Web (BTW 2009), 13. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS)*, Münster, Germany, 2009, Lecture Notes in Informatics. GI.

31.7.3 Selectivity and Cost Estimation

Investigators: Thomas Neumann and Sebastian Michel

The principle goal of query optimization is to find the most efficient query processing strategy for a given query. The query optimizer requires a cost model of the underlying execution engine to estimate the costs of evaluating a certain execution plan, which is used to decide which alternative is better. The most critical parameter of the cost model is the input cardinality of operators, which itself depends on the selectivity of preceding operators. For

example the costs of evaluating the join $\sigma_{p_1}(R_1) \bowtie_{p_2} R_2$ depends primarily on $|\sigma_{p_1}(R_1)|$ and $|R_2|$. The size of base relations ($|R_1|$ and $|R_2|$ in the example) are usually known, but the size of the intermediate result $\sigma_{p_1}(R_1)$ depends both on $|R_1|$ and the selectivity of p_1. The standard method of estimating the selectivity of a predicate is using histograms to approximate the value distributions [1]. Using this approximation, the predicate cardinality is derived from the expected number of qualifying attribute values.

While this general mechanism is well understood, typical histograms have two main drawbacks: first, they usually provide no error bounds (or only L_2 bounds, which are not very useful in practice), and second, they tend to discretize their prediction such that small perturbations to the selection predicate do not affect the predicted cardinality. This is unfortunate for query optimization tasks for top-k processing, where the query optimizer can manipulate selection predicates to improve query performance. We therefore developed a different kind of histograms, using function approximation: We approximated the cumulative distribution function (CDF) of an attribute using linear splines by fitting splines with a minimum number of base points in error corridors around the CDF [2]. Due to the nature of this construction method, we have guaranteed error bounds over the whole value domain, and get accurate estimates for arbitrary values. This significantly improved the results of our optimization techniques and provided very robust estimates.

References

[1] Y. E. Ioannidis. The history of histograms (abridged). In *VLDB*, 2003, pp. 19–30.

- [2] T. Neumann and S. Michel. Smooth interpolating histograms with error guarantees. In W. A. Gray, K. G. Jeffery, and J. Shao, eds., *Sharing Data, Information and Knowledge, 25th British National Conference on Databases, BNCOD 25*, Cardiff, UK, 2008, *LNCS 5071*, pp. 126–138. Springer.

31.7.4 Distributed Top-k Query Processing

Investigators: Thomas Neumann and Sebastian Michel

The usual goal of query processing is to compute the (complete) query result as efficiently as possible. But for interactive queries that are presented to the user, the goal is usually not to compute the complete result, but only to produce the k most relevant results as quickly as possible. This restriction to the top-k results changes query processing quite significantly. In the centralized setting this is well understood and commonly based upon the family of thresholding algorithms [2], but when the data is distributed over multiple nodes in the network these technique are no longer applicable due to the high costs of network traffic.

Instead, specialized distributed top-k algorithms have been proposed, most notably, the TPUT algorithm [1]. When querying n nodes, the TPUT algorithm first contacts all nodes to get their local top-k results (ordered by a *score* function), then computes the lower bound for the global top-k scores (called *min-k*), and then retrieves from each peer all results that are above $\frac{min-k}{n}$. The querying node then aggregates the results, requests missing partial scores if necessary, and produces the final top-k result. The TPUT algorithm produces the correct results without the large number of round-trips implied by centralized algorithms, but it still has suboptimal runtime behavior: The querying node asks all nodes for all data

items above the threshold, independent of network topology or the data distribution within in the network.

We therefore developed a different execution strategy for distributed top-k queries. We formulated the top-k processing using algebraic operators, which allows for hierarchical aggregation and distribution of the top-k computation [6, 7]. The base top-k operators produce the local top-k lists on a node and produce all data items above the desired threshold; the intermediate top-k operators aggregate partial top-k results and prune out data items that cannot make it into the final result; and the final top-k operator examines the input, lookups up missing partial scores, and produces the final result. By using algebraic equivalences we can find alternative algebraic formulations for a given query, and employ a dynamic programming strategy to construct the most efficient execution strategy given a model of the network topology and statistics about the data distributions. The great advantage of this fine grained scheduling of top-k steps is that the computation can now adapt itself to the data placement. For example, it is often beneficial to send small intermediate results to a node with many candidate data items in order to aggregate (and prune) the data early and thus save network traffic costs. This hierarchical execution model greatly improved execution times.

For queries accessing a large number of nodes or where the data is distributed very unevenly we can improve performance even more by carefully adjusting the thresholds propagated to the nodes [5]. For large queries, the uniform threshold computed by TPUT ($\frac{min-k}{n}$) becomes very small, which means that nodes send most of their data over the network. But thresholds do not have to be chosen uniformly, the only requirement is that (when using summation as aggregation), the sum of all thresholds must be $\leq$ *min-k*. Furthermore we observe that score distributions can vary greatly between nodes, which means that some nodes send data over a long time, while others finish quickly. As we have to wait for the last node to finish before computing the result, we adapt the thresholds for individual nodes such that all nodes produce the same amount of data (or, when using a detailed network model, use the same time period). This results in a much more balanced execution of top-k queries, and again improved execution time. This is also an area where our work on selectivity estimation is crucial (see Section 31.7.3), as the threshold tuning algorithms needs very accurate models of the score distributions within the nodes to compute good thresholds.

An interesting variation of this theme are queries with decreasing aggregation functions. Typical examples are *min* or set intersection, which are commonly used during node selection in distributed information retrieval, for example when selecting most relevant peers for multi-keyword queries. Here the standard techniques for computing score bounds return meaningless bounds, which causes all nodes to send all their data over the network. We addressed this issue by piggybacking probes and bitmap filters into the regular retrieval process, improving bounds as quickly as possible during query execution [3]. In the centralized setting we also studied the problem of ranked retrieval, i.e., operating on sorted data without explicit scoring function, but with more complex operators. We proposed using rank-aware algebraic operators [4] that predict how much input they will require to compute their individual output, and use batched block operations to compute additional input in case their prediction is wrong. This greatly improves query execution times and supports a wide range of queries including selection-projection-join and group-by queries.

References

[1] P. Cao and Z. Wang. Efficient top-k query calculation in distributed networks. In *PODC*, 2004, pp. 206–215.

[2] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. In *PODS*, 2001. ACM.

- [3] S. Michel and T. Neumann. Search for the best but expect the worst - distributed top-k queries over decreasing aggregated scores. In *Proceedings of the Tenth International Workshop on the Web and Databases (WebDB 2007)*, Beijing, China, 2007. ACM.
- [4] T. Neumann. Optimizing ranked retrieval. In W. Wagner, N. Revell, and G. Pernul, eds., *18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, Regensburg, Germany, 2007, *LNCS 4653*, pp. 329–338. Springer.
- [5] T. Neumann, M. Bender, S. Michel, R. Schenkel, P. Triantafillou, and G. Weikum. Optimizing distributed top-k queries. In *Proceedings of the 9th International Conference on Web Information Systems (WISE 2008)*, Auckland, New Zealand, 2008, *LNCS 5175*, pp. 337–349. Springer.
- [6] T. Neumann and S. Michel. Algebraic query optimization for distributed top-k queries. In A. Kemper, H. Schöning, T. Rose, M. Jarke, T. Seidl, C. Quix, and C. Brochhaus, eds., *12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web*, Aachen, Germany, 2007, *Lecture Notes in Informatics*, vol. 103, pp. 324–343. Gesellschaft für Informatik.
- [7] T. Neumann and S. Michel. Algebraic query optimization for distributed top-k queries. *Informatik - Forschung und Entwicklung*, 21(3-4):197–211, 2007.

31.8 Distributed Data and Communities

Coordinators: Christos Tryfonopoulos (until June 2008) *and Mauro Sozio* (since July 2008)

Popular Web 2.0 applications and rapidly increasing social online communities face the need for distributed data management for scalability reasons. They may serve millions of users and manage user-provided data that is distributed across many computers in a data center or even over a wide-area network. To ensure good search performance and high availability of services, such distributed systems must have mechanisms and intelligent strategies to cope with high failure rates of components and with high dynamics of data, workload, and the network itself. Peer-to-peer networks for data sharing pose similar challenges and additionally face the problem of misbehaving peers that aim to manipulate distributed computations to their advantage. This entails the need for modeling and computing authority, trust, and reputation measures to retrieve relevant content and identify trustworthy peers.

31.8.1 Dynamic Replication in Peer-to-Peer Networks

Investigators: Mauro Sozio, Thomas Neumann, Stefan Holder, and Gerhard Weikum

Replication is widely used in distributed data management for several reasons. First, replicating data items at different nodes in the network improves reliability (i.e., the probability of not losing any data even in the presence of multiple permanent node failures) and availability (i.e., the probability of having the data accessible even in the presence of multiple temporary node outages). Second, having a choice among multiple replicas might lead to better load balance,

higher throughput as well as shorter response times. Third, search requests in large-scale networks are often processed by exploring only a limited fraction of the network, allowing any request to fail with some probability. In this scenario, replication improves the *probability of successfully retrieving* the requested item. A canonical example for this case are *unstructured peer-to-peer (P2P) overlay networks* such as Gnutella or BitTorrent with file sharing among large user communities.

The protocols for P2P replication management and processing search requests are fairly mature [4]. What is much less understood are strategies for the creation and placement of replicas in such large-scale, heterogeneous, highly dynamic networks. Two key questions are: how many replicas should a data item have, and where in the network do we place these. Good solutions to these questions need to consider the request rate (popularity) of each data item, the distribution of query originators for the requests to an item, the network topology and per-link performance characteristics, the storage capacity of each peer, and many more parameters. What makes the problem even harder, is that the replica placement has to be computed in a decentralized setting. This entails that such a global objective has to be achieved despite the fact that each peer has initially a local knowledge of the network (usually restricted to its immediate neighbors) and can exchange only a limited amount of messages. Not surprisingly, the solutions proposed in the literature are based on much simpler computational models that takes into account only a subset of these characteristics of realistic systems.

One of the most elegant and widely recognized contributions aiming at the above questions is the work of Cohen and Shenker [1]. This paper aims to minimize the runtime needed to successfully retrieve a data item in the peer-to-peer network. The search protocol considered in [1] is rather simple (and perhaps unrealistic): the peer looking for a data item samples peers from the network, independently at random, until the requested data item has been found. The main result of the paper is a closed formula for the optimal number of replicas to minimize the expected number of sampled peers.

Notwithstanding the depth and elegance of this work, the proposed solution presents some practical issues. Above all, the search protocol considered is not realistic as reaching peers far away in the network might be expensive. Another limitation, is that some "global" parameters need to be known in advance by peers, making a distributed implementation non-obvious. Coping with dynamic network parameters seems to be even harder. Another interesting result is the work of Morselli et al. [3]. Unfortunately, the nice theoretical guarantees shown in this paper, hold solely in the case when the network topology has some specific properties, while in some cases of interest the proposed solution degrades in efficiency.

In [5] we propose a completely decentralized algorithm, coined P2R2 (P2P Replication with 2-Approximation), addressing all the unsolved issues of the previous approaches. We consider a search protocol, where each query initiator probes a bounded neighborhood by means of length-bounded random walks. In fact, this is a quite realistic abstraction of the bounded flooding schemes adopted by real peer-to-peer file-sharing systems. In our settings, peers have a bounded capacity and data items have a given size. Moreover, we also handle the case when data items have different popularity (query rate) in different peers. We prove that the P2R2 algorithm can guarantee a successful-search probability that is within a factor of 1/2 of the optimal solution. Moreover, P2R2 automatically adjusts to changes in the network, achieving a 1/2-approximation guarantee every time the network parameters are stable for

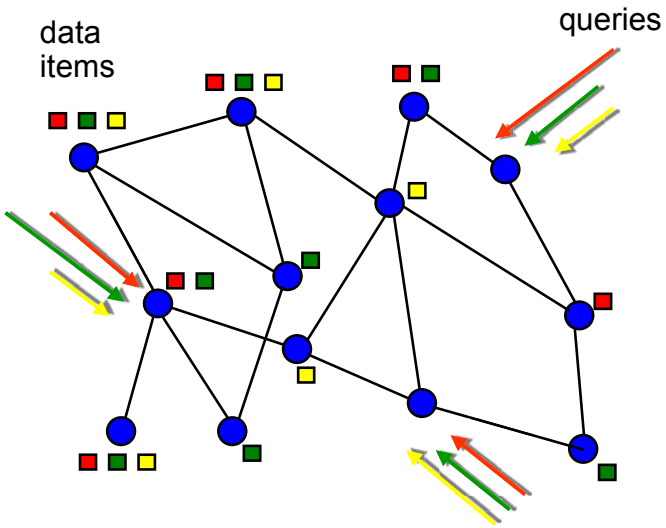


Figure 31.12: A P2P network with a set of data items where peers issue queries at different rates.

"sufficiently long" time. These theoretical studies are complemented with experimental results showing the effectiveness and the efficiency of our approach in comparison with the existing literature. Figure 31.12 illustrates a P2P network together with a set of data items (the colorful rectangles) where peers issue queries at different rates (query rates are represented as arrows having different length).

Peer-to-peer networks are usually very dynamic, as peers might join or leave the network at any time. They are also characterized by frequent temporary failures of peers, caused by crashes in their machines or simply by terminating the file-sharing software. Under these circumstances, replicas located at the faulty peer might be unavailable for some time. Nevertheless, we wish to ensure that data items in the network can be retrieved with some guarantee. A first contribution of [2] is a formal definition of this problem as follows. For each peer we are given an online probability indicating the probability for this peer of being available at time t. In addition, each peer i specifies for each item j a rational number a_{ij} in the range between zero and one requesting that any query for item j should be successful with probability at least a_{ij}. The search protocol considered is a length-bounded random walk. The challenge is then how to place replicas in the network in such a way all that availability guarantees are met and the number of replicas is minimized.

We first study this problem in a centralized setting, devising an $O(\log n + \log m)$-approximation algorithm where n and m are the number of peers and the number of items, respectively. Then, we adapt the algorithm to a distributed setting. Our theoretical analysis is complemented with extensive experiments measuring the number of replicas placed as a function of the online probabilities, the availability guarantees as well as the length of the random walk. Our experiments show the practical viability of our approach.

References

[1] E. Cohen and S. Shenker. Replication strategies in unstructured peer-to-peer networks. In *SIGCOMM*, 2002, pp. 177–190. ACM.

- [2] S. Holder. Replication in unstructured peer-to-peer networks with availability constraints. Masters thesis, Universität des Saarlandes, 2008.

[3] R. Morselli, B. Bhattacharjee, M. A. Marsh, and A. Srinivasan. Efficient lookup on unstructured topologies. *IEEE Journal on Selected Areas in Communications*, 25(1):62–72, 2007.

[4] Y. Saito and M. Shapiro. Optimistic replication. *ACM Comput. Surv.*, 37(1):42–81, 2005.

- [5] M. Sozio, T. Neumann, and G. Weikum. Near-optimal dynamic replication in unstructured peer-to-peer networks. In M. Lenzerini and D. Lembo, eds., *Proceedings of the Conference on Principles of Database System (PODS 2008)*, Vancouver, Canada, 2008, pp. 281–290. ACM.

31.8.2 Distributed Link Analysis

Investigators: Josiane Parreira, Sebastian Michel, and Gerhard Weikum

Analyzing the authority or reputation of entities that are connected by a graph structure and ranking these entities is an important issue that arises in the Web, in Web 2.0 communities, and in other applications. As Google has impressively demonstrated with its PageRank algorithm, such authority information can be exploited for improving the rank of search results. Although such analyses could be performed by a centralized server, there are good reasons that suggest running theses computations in a decentralized manner across many peers, like scalability, privacy, censorship, etc. Previous distributed approaches are not suitable for a peer-to-peer network, where overlap among the fragments usually occur. In addition, peer-to-peer approaches also need to consider network characteristics, such as peers unaware of other peers' contents, susceptibility to malicious attacks, and network dynamics (so-called churn).

JXP [1, 2] is a decentralized algorithm for computing authority scores of entities distributed in a peer-to-peer (P2P) network that allows peers to have overlapping content and requires no a priori knowledge of other peers' content. It combines independent PageRank computation on each peer's local portion of the graph with equally independent meetings among peers in the network, for knowledge exchange.

In recent work [4], we have addressed the issue of network dynamics: peers are constantly joining and leaving the network, meaning that their content is not always available. Moreover, peers might change what they store, for instance a user can become interested in a different topic and start to store information about this new topic instead. Network dynamics has a big impact on the computation of authority scores, since the endorsement links might as well change. We have developed a distributed algorithm that estimates the number of distinct entities in the network, which is needed in the local computation of the PageRank scores. The algorithm provides an efficient estimation, with little overhead on the network traffic. We have also extended the JXP algorithm for coping with the network dynamics [4].

In the context of social networks, we have recently demonstrated how JXP scores can be used to rank entities in online communities [3].

References

- [1] J. X. Parreira, C. Castillo, D. Donato, S. Michel, and G. Weikum. The Juxtaposed approximate PageRank method for robust PageRank approximation in a peer-to-peer Web search network. *VLDB Journal*, 17(2):291–313, 2008.

  [2] J. X. Parreira, D. Donato, S. Michel, and G. Weikum. Efficient and decentralized pagerank approximation in a peer-to-peer web search network. In U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim, eds., *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB 2006)*, Seoul, Korea, 2006, pp. 415–426. ACM. Acceptance ratio 1:7.

- [3] J. X. Parreira, S. Michel, M. Bender, T. Crecelius, and G. Weikum. P2P authority analysis for social communities. In C. Koch, J. Gehrke, M. N. Garofalakis, D. Srivastava, K. Aberer, A. Deshpande, D. Florescu, C. Y. Chan, V. Ganti, C.-C. Kanne, W. Klas, and E. J. Neuhold, eds., *33rd International Conference on Very Large Data Bases (VLDB 2007)*, Vienna, Austria, 2007, pp. 1398–1401. ACM.

- [4] J. Xavier Parreira, S. Michel, and G. Weikum. Efficiently handling dynamics in distributed link based authority analysis. In J. Bailey, D. Maier, K.-D. Schewe, B. Thalheim, and X. S. Wang, eds., *9th International Conference on Web Information Systems Engineering (WISE)*, Auckland, New Zealand, 2008, pp. 36–49. Springer.

31.8.3 Trust and Misbehavior in Peer-to-Peer Communities

Investigators: Tom Crecelius and Mauro Sozio

Peer-to-Peer (P2P) systems provide advantages over centralized systems due to their scalability and computational power. Moreover, regarding reliability, privacy and autonomy issues, P2P systems are less vulnerable than centralized settings, as users participating in P2P communities may keep their contents locally on their own system, rather than giving contents away to a centralized Web service and relying upon a single source of failure (including the possibility of insider misuse or sabotage).

In order to enable users of a P2P community to explore and search new content by simply submitting queries to the P2P network, peers have to collaborate and form a distributed search engine. As each user's peer is a part of the search engine, this opens a door for misbehavior. Users may try to cheat by violating the protocol of collaboration for various reasons, e.g., to deceptively influence their own reputation in a P2P community. In our research work [8], we address the problem of countering misbehavior in authority computations over social networks in a P2P environment.

Eigenvector computations are an important building block for computing authority, trust, and reputation scores in social networks and other graphs and, in fact, distributed algorithms for Eigenvector computations using the Jacobi power iteration method [4] and other spectral analyses have received quite some attention in prior research, most notably for but not limited to Web-graph link analysis [1, 2, 3, 5, 6, 7, 9, 10].

In decentralized settings, like P2P communities, this kind of analysis requires bilateral data exchanges between peers, and therefore, gives rise to the problem that dishonest peers may cheat in order to manipulate the computation's outcome. To counter cheating, our solution is based on general principles of replication and randomization and thus widely applicable to social network analysis, Web link analysis, and other problems of this kind. Hence, it works

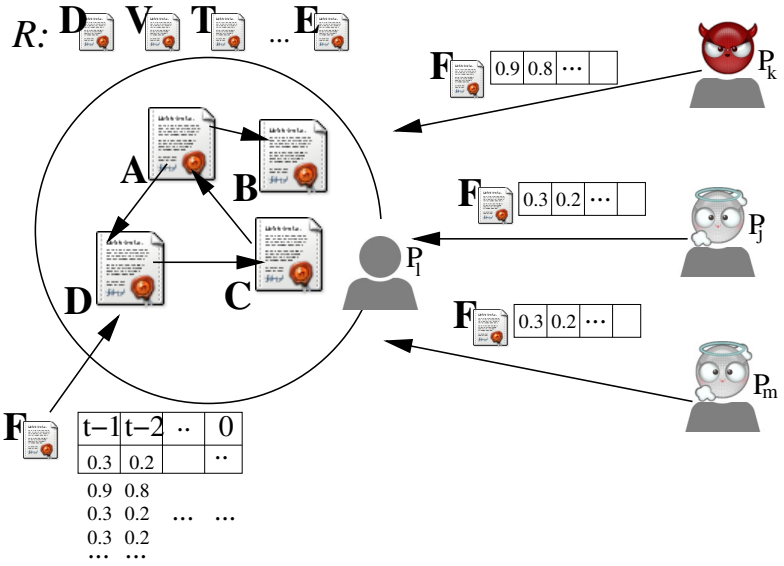


Figure 31.13: Combating cheating in a P2P environment

in a perfectly asynchronous manner, can handle arbitrary distributions of the data across peers, and it converges to the correct result that the honest peers alone would compute.

Our algorithm is based on two key principles. First, we introduce a bounded amount of redundancy into the peer network by replicating the entities of interest (e.g., Web pages or users of a social network) in a randomized manner and, second, whenever two peers exchange information about the authority of an entity, they provide a version history of the entity's previous values, thus enabling the receiving peers to compare values from different peers or at different times in a meaningful manner.

Figure 31.13 sketches the idea of our algorithm. Each peer is responsible for a randomized subset R of all replicated entities and only computes authority scores for these entities. By meeting other peers, it gathers authority information to improve its own computations but also accepts information only if a peer is responsible for the associated entity. Each step t of the power iteration method is based on the previous step $t-1$. During random meetings of a peer P_l with other peers (e.g., P_k is malicious, P_m and P_j are honest peers), score values of all iteration steps that already have been computed for an entity of interest F have to be exchanged. Moreover, peer P_l keeps a version history for all values received and determines the majority of votes m for each iteration step.

Our algorithm works under the realistic assumptions that the fraction of cheating peers is only a minority and that peer identities are unforgeable; the latter one can be ensured by using standard methods of cryptographic security. We prove that our algorithm computes the correct authority values, as stated by the following theorem.

Theorem 1 (Convergence) *Consider a peer-to-peer network with n peers, p n of which are malicious ($p < 1/2$). Let j be an entity held by peer i and let $\pi^t(j)$ be the authory score of*

j computed by the (centralized) power-iteration method at step t. After $4\,t\,n^2 \log n$ random meetings, the authority score of j computed by i matches $\pi^t(j)$, for all i, j, t, with high probability.

Additionally, we prove a lower bound for the communication complexity in the simple case where there are only two parties involved in the computation. This highlights the inherent trade-off between the convergence speed and the messaging cost.

Theorem 2 (Lower bound) *In the two-party model, any deterministic protocol computing a $\frac{1}{\sqrt{2-\alpha}}$-approximation of the top-entity score requires $\Omega(\frac{m}{\log^2 m})$ bits to be sent, where m is the total number of entities and α is the damping factor in the PageRank algorithm.*

Furthermore, we provide experimental evidence, based on an excerpt of a social-tagging community, that the algorithm is practically viable, in terms of the necessary degree of replication, convergence speed, and communication overhead.

References

[1] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *WWW*, 2003, pp. 280–290.

[2] J. V. Davis and I. S. Dhillon. Estimating the global PageRank of Web communities. In *KDD*, 2006, pp. 116–125.

[3] D. Kempe and F. McSherry. A decentralized algorithm for spectral analysis. In *STOC*, 2004, pp. 561–568.

[4] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings.* Princeton University Press, Princeton, NJ, USA, 2006.

- [5] J. X. Parreira, C. Castillo, D. Donato, S. Michel, and G. Weikum. The Juxtaposed approximate PageRank method for robust PageRank approximation in a peer-to-peer Web search network. *VLDB Journal*, 17(2):291–313, 2008.

[6] K. Sankaralingam, M. Yalamanchi, S. Sethumadhavan, and J. C. Browne. PageRank computation and keyword search on distributed systems and P2P networks. *J. Grid Comput.*, 1(3):291–307, 2003.

[7] S. Shi, J. Yu, G. Yang, and D. Wang. Distributed page ranking in structured P2P networks. In *ICPP*, 2003, pp. 179–186. IEEE Computer Society.

- [8] M. Sozio, T. Crecelius, J. Xavier Parreira, and G. Weikum. Good guys vs. bad guys: Countering cheating in peer-to-peer authority computations over social networks. In *11th International Workshop on the Web and Databases (WebDB 2008)*, Vancouver, Canada, 2008, pp. 103–108. ACM.

[9] Y. Wang and D. J. DeWitt. Computing PageRank in a distributed Internet search engine system. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, eds., *VLDB*, 2004, pp. 420–431. Morgan Kaufmann.

[10] J. Wu and K. Aberer. Using a layered Markov model for distributed Web ranking computation. In *ICDCS*, 2005, pp. 533–542. IEEE Computer Society.

31.8.4 Peer-to-Peer Query Routing

Investigators: Srikanta Bedathur, Matthias Bender, Sebastian Michel, Paraskevi Raftopoulos, Christos Tryfonopoulos, and Christian Zimmer

Document collections nowadays are fragmented across different Digital Libraries (DLs) due to copyright issues that prevent the owners to share their documents. To deal with this issue, a number of P2P architectures (most using a Distributed Hash Table - DHT [9, 10] as the underlying routing infrastructure) that allow users to transparently search these data collections [4, 11, 12, 15] have emerged as a natural decentralized solution. In this line of work, we introduce iClusterDL [6], a novel P2P architecture utilizing self-organizing Semantic Overlay Networks (SONs) [1, 5], that demonstrates (i) the feasibility of supporting rich query models without resorting to structured overlays, and (ii) the benefits derived from a looser organization of system components. iClusterDL relies on the idea of organizing peers into SONs, where clusters are linked by virtue of containing similar information. Building upon a P2P model, iClusterDL consists of three general types of peers, namely information providers (contributing documents to the network), information consumers (seeking for existing information), and super-peers. Super-peers act as access points for clients and providers. They self-organize into a SON to offer a robust, fault-tolerant and scalable means for routing messages and managing queries. The description of each super-peer is derived from the descriptions of the providers connected to it, allowing super-peers to organize into clusters of similar content in the spirit of [7, 8].

iClusterDL is designed to support both information retrieval (IR) and information filtering (IF) functionality. In an IR scenario a user poses a one-time query and the system returns all resources matching the query (e.g., all currently available documents relevant to the query). In an IF scenario a user submits a continuous query and waits to be notified about certain future events of interest (i.e., about newly published documents relevant to the continuous query). iClusterDL offers an infrastructure, based on concepts of P2P systems, for organizing the super-peers in a scalable, efficient and self-organizing architecture. This architecture (i) allows for seamless integration of information sources, since different DLs and other content providers offer the same querying interface through the iClusterDL system to the end-user, (ii) enhances fault tolerance, since it avoids centralized components that introduce bottlenecks and single points of failure, and (iii) requires no central administration authority, since each participant is responsible for the administration and maintenance of its own (super-)peers.

iClusterDL is the first approach towards efficient organization of DLs in SONs that supports both IR and IF functionality. The proposed architecture is automatic (requires no intervention and minimal administration), general (requires no previous knowledge of the DL contents and works for any type of data model or content), adaptive (adjusts to changes of DL contents), efficient (offers fast query processing) and accurate (achieves high recall).

State-of-the-art Peer-to-Peer Information Retrieval (P2P IR) systems suffer from their lack of response time guarantee especially with scale. To address this issue, a number of techniques for caching of multi-term inverted list intersections and query results have been proposed recently. Although these enable speedy query evaluations with low network overheads, they fail to consider the potential impact of caching on result quality improvements. In [13, 14] we propose the use of a cache-aware query routing scheme that not only reduces the response

delays for a query, but also presents an opportunity to improve the result quality while keeping the network usage low. In this regard, there are three-fold contributions in this paper: First of all, a cache-aware, multi-round query routing strategy is developed that balances between query efficiency and result-quality. This Exact Caching (EC) approach stores cached results in the directory in a smart manner such that caching is integrated in the standard query execution process of our P2P search engine MINERVA [2, 3, 15]. The querying peer can increase the query result-quality by asking additional peers not involved in the previous result. The large-scale experiments used a Wikipedia benchmark and a real-world query-log. Various cache replacement strategies are investigated and churn in the P2P system is considered. Next, the paper proposes to aggressively reuse the cached results of even subsets of a query towards an approximate caching technique that can drastically reduce the bandwidth overheads, and studies the conditions under which such a scheme can retain good result-quality. If the P2P directory can not deliver information about the exact query but cached results for subqueries, the Approximate Caching (AC) approach returns satisfying outcomes. Nevertheless, to reach a good approximation, the paper makes demands on the existing cached results. In a real world experiment the paper used a large query-log with the combination of our EC and AC approach yielding to satisfying results in terms of relative recall and message costs. Finally, we empirically evaluate these techniques over a fully functional P2P IR system, using a large-scale Wikipedia benchmark, and using both synthetic and real-world query workloads. Our results show that our proposal to combine result caching with multi-round, cache-aware query routing can reduce network traffic by more than half while doubling the result quality.

References

[1] K. Aberer, P. Cudré-Mauroux, M. Hauswirth, and T. V. Pelt. GridVine: Building Internet-scale semantic overlay networks. In *Proceedings of the International Semantic Web Conference (ISWC)*, 2004.

• [2] M. Bender, T. Crecelius, S. Michel, and J. X. Parreira. P2P Web search: Make it light, make it fly (demo). In *3rd Biennial Conference on Innovative Data System Research (CIDR 2007)*, Asilomar, USA, 2007, pp. 164–168. www.crdrdb.org.

[3] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Improving collection selection with overlap-awareness. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, eds., *SIGIR 2005, Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brazil, 2005, pp. 67–74. ACM.

[4] C. Doulkeridis, K. Noervaag, and M. Vazirgiannis. Scalable semantic overlay generation for P2P-based digital libraries. In *European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*, 2006.

[5] A. Loser, M. Wolpers, W. Siberski, and W. Nejdl. Semantic overlay clusters within super-peer networks. In *DBISP2P*, 2003.

• [6] P. Raftopoulou, E. Petrakis, C. Tryfonopoulos, and G. Weikum. Information retrieval and filtering over self-organising digital libraries. In D. Christensen-Dalsgaard, Birte Castelli, B. A. Jurik, and J. Lippincott, eds., *12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*, Aarhus, Denmark, 2008, pp. 320–333. Springer.

[7] P. Raftopoulou and E. G. Petrakis. A measure for cluster cohesion in semantic overlay networks. In *Proceedings of the 6th International Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR)*, Napa Valley, California, 2008. ACM.

[8] P. Raftopoulou, E. G. Petrakis, C. Tryfonopoulos, and G. Weikum. Information Retrieval and Filtering over Self-Organising Digital Libraries. In *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Aarhus, Denmark, 2008. ACM.

[9] A. I. T. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In R. Guerraoui, ed., *Middleware*, 2001, *LNCS 2218*, pp. 329–350. Springer.

[10] I. Stoica, R. Morris, D. R. Karger, M. F. Kaashoek, and H. Balakrishnan. CHORD: A scalable peer-to-peer lookup service for Internet applications. In *SIGCOMM*, 2001, pp. 149–160.

[11] J. Stribling, I. G. Councill, J. Li, M. F. Kaashoek, D. R. Karger, R. Morris, and S. Shenker. OverCite: A cooperative digital research library. In M. Castro and R. van Renesse, eds., *Peer-to-Peer Systems IV, 4th International Workshop, IPTPS 2005, Ithaca, NY, USA, February 24-25, 2005, Revised Selected Papers*, 2005, *LNCS 3640*, pp. 69–79. Springer.

[12] C. Tryfonopoulos, S. Idreos, and M. Koubarakis. LibraRing: An architecture for distributed digital libraries based on DHTs. In A. Rauber, S. Christodoulakis, and A. M. Tjoa, eds., *Research and Advanced Technology for Digital Libraries, 9th European Conference, ECDL 2005, Vienna, Austria, September 18-23, 2005, Proceedings*, 2005, *LNCS 3652*, pp. 25–36. Springer.

- [13] C. Zimmer, S. Bedathur, and G. Weikum. Standing on the shoulders of peers: Caching in peer-to-peer information retrieval. In *Fifth International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P)*, Vienna, Austria, 2007.

- [14] C. Zimmer, S. Bedathur, and G. Weikum. Flood little, cache more: Effective result-reuse in P2P IR systems. In J. R. Haritsa, K. Ramamohanarao, and V. Pudi, eds., *13th International Conference on Database Systems for Advanced Applications (DASFAA)*, New Delhi, India, 2008, *LNCS 4947*, pp. 235–250. Spinger.

- [15] C. Zimmer, C. Tryfonopoulos, and G. Weikum. MinervaDL: An architecture for information retrieval and filtering in distributed digital libraries. In L. Kovács, N. Fuhr, and C. Meghini, eds., *Research and Advanced Technology for Digital Libraries, 11th European Conference, ECDL 2007*, Budapest, Hungary, 2007, *LNCS 4675*, pp. 148–160. Springer.

31.8.5 Peer-to-Peer Publish-subscribe Services

Investigators: Christos Tryfonopoulos and Christian Zimmer

Today's content providers are naturally distributed and produce large amounts of new information every day, making peer-to-peer (P2P) data management a promising approach that offers scalability, adaptivity to high dynamics, and failure resilience. One-time querying in such a P2P setting is unarguably the most popular user activity, however subscribing with a continuous query (also referred to as information filtering, publish/subscribe or alerting) is of equal importance as it allows the user to cope with the high rate of information production and avoid the cognitive overload of repeated searches. Subscribing users, or services that act on users' behalf, specify continuous queries, thus expecting to be notified when newly appearing documents that satisfy the query conditions are made available in the system.

All approaches to P2P data management taken so far, focus either on *exact* information retrieval (IR) [2, 4] or information filtering (IF) [1, 5] by using the P2P network as a decentralized index for both documents and continuous queries. To facilitate this indexing, appropriate protocols that disseminate documents and queries in a deterministic way, depending on the terms contained in them are employed. These document and query indexing protocols lead to filtering effectiveness that is exactly the same as that of a centralized system. This, however, creates an efficiency and scalability bottleneck, while in certain applications this design might not even be desirable (e.g., in applications like news of blog filtering, where the user is not interested in all relevant items, but rather in the most interesting ones).

Contrary to approaches that provide exact IR and IF functionality by utilizing per-document indexing, in MAPS [7, 9, 11, 12] the concept of approximate IR and IF is introduced; publications are processed locally and peers query or subscribe to only a few, selected information sources that are most likely to satisfy the user's information demand. In this way, per-peer (rather than per-document) indexing is employed and efficiency and scalability are enhanced by trading a small reduction in recall for lower message traffic.

The MAPS system utilizes a structured overlay to support publisher selection and ranking necessary for both IR and IF scenarios. This selection is driven by statistical summaries stored in a distributed P2P directory built on top of the Pastry DHT [3]. For scalability, summaries have publisher and not document granularity, thus capturing the best publisher for certain keywords but not for specific documents. Both approximate IR and IF services utilize the same conceptually global, but physically distributed directory of statistical metadata to derive information provider rankings. To support the IR functionality, MAPS utilizes well-known resource selection techniques for P2P query routing such as TF/IDF based methods, CORI, or language models to route the user query to a carefully selected subset of information sources. Resource selection in such an autonomous and dynamic environment can be improved by taking into account the overlap in the document collections of different content providers [2].

To support P2P IF in a scalable and efficient way, MAPS ranks sources, and delivers matches only from the best ones, by utilizing novel publisher selection strategies. Thus, the continuous query is replicated to the best information sources and only published documents from these sources are forwarded to the subscriber. This approximate IF relaxes the assumption, which holds in most IF systems, of potentially delivering notifications from every producer and amplifies scalability [6, 8, 10]. To select the most appropriate publishers to subscribe to, a subscriber computes scores that reflect the past publishing behavior and utilizes them to predict future peer behavior. This score is based on a combination of resource selection (i.e., TF/IDF based) and behavior prediction to deal with the dynamics of publishing [6, 8]. Behavior prediction uses time-series analysis with double exponential smoothing techniques to predict future publishing behavior, and adapt faster to changes in it. In addition, correlations among keywords in multi-term continuous queries can be exploited to further improve publisher selection. In [12], two such strategies based on statistical synopses are described in detail. In this way, approximate IF achieves higher scalability by trading faster response times and lower message traffic for a moderate loss in recall.

References

[1] I. Aekaterinidis and P. Triantafillou. PastryStrings: A comprehensive content-based publish/subscribe DHT network. In *26th IEEE International Conference on Distributed Computing Systems (ICDCS 2006), 4-7 July 2006, Lisboa, Portugal*, 2006, p. 23. IEEE Computer Society.

[2] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Improving collection selection with overlap-awareness. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, eds., *SIGIR 2005, Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brazil, 2005, pp. 67–74. ACM.

[3] A. I. T. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In R. Guerraoui, ed., *Middleware*, 2001, *LNCS 2218*, pp. 329–350. Springer.

[4] O. Sahin, F. Emekci, D. Agrawal, and A. Abbadi. Content-based similarity search over peer-to-peer systems. In *Proceedings of the International Workshop on Databases, Information Systems, and Peer-to-Peer Computing (DBISP2P)*, 2004.

[5] C. Tryfonopoulos, S. Idreos, and M. Koubarakis. Publish/subscribe functionality in IR environments using structured overlay networks. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, eds., *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, 2005, pp. 322–329. ACM.

- [6] C. Tryfonopoulos, C. Zimmer, M. Koubarakis, and G. Weikum. Architectural alternatives for information filtering in structured overlay networks. *IEEE Internet Computing*, 11(4):24–34, 2007.

- [7] C. Zimmer, J. Heinz, C. Tryfonopoulos, and G. Weikum. P2P information retrieval and filtering with MAPS (demo). In K. Wehrle, W. Kellerer, S. K. Singhal, and R. Steinmetz, eds., *Proceedings of the Eighth International Conference on Peer-to-Peer Computing (P2P 2008)*, Aachen, Germany, 2008, pp. 84–85. IEEE Computer Society.

- [8] C. Zimmer, C. Tryfonopoulos, K. Berberich, M. Koubarakis, and G. Weikum. Node behavior prediction for large-scale approximate information filtering. In *1st International Workshop on Large Scale Distributed Systems for Information Retrieval (LSDS-IR 2007)*, Amsterdam, The Netherlands, 2007.

- [9] C. Zimmer, C. Tryfonopoulos, K. Berberich, M. Koubarakis, and G. Weikum. Approximate information filtering in peer-to-peer networks. In J. Bailey, D. Maier, K.-D. Schewe, B. Thalheim, and X. Sean Wang, eds., *Proceedings of the 9th International Conference on Web Information Systems Engineering (WISE 2008)*, Auckland, New Zealand, 2008, *LNCS 5175*, pp. 6–19. Springer.

- [10] C. Zimmer, C. Tryfonopoulos, and G. Weikum. Efficient search and approximate information filtering in a distributed peer-to-peer environment of digital libraries. In C. Thanos, F. Borri, and L. Candela, eds., *Digital Libraries: Research and Development, First International DELOS Conference*, Pisa, Italy, 2007, *LNCS 4877*, pp. 328–337. Springer.

- [11] C. Zimmer, C. Tryfonopoulos, and G. Weikum. MinervaDL: An architecture for information retrieval and filtering in distributed digital libraries. In L. Kovács, N. Fuhr, and C. Meghini, eds., *Research and Advanced Technology for Digital Libraries, 11th European Conference, ECDL 2007*, Budapest, Hungary, 2007, *LNCS 4675*, pp. 148–160. Springer.

• [12] C. Zimmer, C. Tryfonopoulos, and G. Weikum. Exploiting correlated keywords to improve approximate information filtering. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, eds., *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, 2008, pp. 323–330. ACM.

31.8.6 Anonymous and Censorship-resilient Data Sharing

Investigators: Christos Tryfonopoulos and Christian Zimmer in cooperation with Michael Backes, Marek Hamerlik, Alessandro Linari, and Matteo Maffei

Over the last years unstructured overlays have evolved as a natural decentralized way to share data and services among a network of loosely connected peers. In unstructured overlays, peers connect to a small set of other peers, and queries are propagated along overlay connections, using some query forwarding strategy that aims at finding peers with resources matching the issued query. Semantic Overlay Networks (SONs) [2, 4, 5] are an instance of unstructured overlays, where peers that are semantically, thematically, or socially close are organized into groups to exploit similarities at query time. This flexible organization maintains high peer autonomy, and has proved a useful technology not only for distributed information retrieval, but also as a natural distributed alternative to Web 2.0 application domains such as decentralized social networking in the spirit of Flickr or del.icio.us.

In such an information exchange setting, not all information providers are willing to reveal their true identity: for instance, publishers may want to present their opinions anonymously to avoid associations with their race, ethnic background or other sensitive characteristics. Furthermore, people seeking for sensitive information may want to remain anonymous so as to avoid being stigmatized. The freedom of information exchange is another important issue that got increasing attention in the last years, since organizations, such as governments or private companies, may regard a discussion topic or a report as inconvenient or harmful and try to censor it.

To address these issues, we put forward Clouds [3], a novel P2P search infrastructure for providing *anonymous* and *censorship resistant* search functionality in a SON. We achieve anonymity by relying on a self-organization of peers into groups that we call *clouds*. Message routing is modified to take place among clouds instead of peers, thus hiding the identity of both the resource provider and the querying peer, while cloud size is a tunable parameter that affects anonymity and efficiency. Censorship resistance at communication level is achieved by a cryptographic protocol that guarantees the secrecy of the resource, thus avoiding censorship based on the inspection of the messages circulating in the network. The design of such a protocol needs to meet a number of challenging goals: allowing for the exchange of encrypted messages without assuming previously shared secrets, avoiding centralized infrastructures, like trusted servers or static gateways, and guaranteeing efficiency without establishing direct connections between peers. Clouds is the first system to guarantee anonymity *and* censorship resistance in SONs. It extends the notion of clouds proposed in [1, 6] to achieve anonymity in a SON setting, while considering *dynamic cloud creation* and avoiding *static cloud gateways*. Additionally it utilizes a novel cryptographic protocol for securing the communication channel between participants and achieving censorship resistance.

References

[1] A. Singh and B. Gedik and L. Liu. Agyaat: Mutual anonymity over structured P2P networks. *Emerald Internet Research Journal*, 2006.

[2] K. Aberer, P. Cudré-Mauroux, M. Hauswirth, and T. V. Pelt. GridVine: Building Internet-scale semantic overlay networks. In *Proceedings of the International Semantic Web Conference (ISWC)*, 2004.

- [3] M. Backes, M. Hamerlik, A. Linari, M. Maffei, C. Tryfonopoulos, and G. Weikum. Anonymous and censorship resistant content sharing in unstructured overlays. In R. A. Bazzi and B. Patt-Shamir, eds., *Twenty-Seventh Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2008)*, Toronto, Canada, 2008, pp. 429–429. ACM.

[4] A. Crespo and H. Garcia-Molina. Semantic overlay networks for P2P systems. Technical Report, Stanford University, 2003.

[5] A. Loser, M. Wolpers, W. Siberski, and W. Nejdl. Semantic overlay clusters within super-peer networks. In *DBISP2P*, 2003.

[6] M. K. Reiter and A. D. Rubin. Crowds: anonymity for Web transactions. *ACM Trans. Inf. Syst. Secur.*, 1(1):66–92, 1998.

31.8.7 Distributed Statistics Management

Investigators: Gerhard Weikum in cooperation with Nikos Ntarmos and Peter Triantafillou

Internet-scale applications have a need for distributed aggregation queries over a large number of network nodes which can be dynamically selected by filter predicates. In many cases, the aggregates have a statistical nature, and can thus sometimes be estimated with sufficient accuracy and without computing the full, exact result [1]. For example, in peer-to-peer file sharing, one may be interested in estimating the total number of distinct files that the entire network has available at a given point – a CountDistinct aggregation. In e-science, the total number of X-ray spectras for triple star systems available anywhere in a Grid network could be an interesting measure, requiring a distributed Count operation. For the analysis of Internet-traffic logs, the total amount of bytes transferred to clients in a particular range of IP addresses can be computed by a distributed Sum operation. Last but not least, the advanced join-and-group queries over structured data sources typically require choosing a low-cost query execution plan, and this query optimization in turn relies on sufficiently accurate statistical estimators for the selectivity of query predicates (i.e., the cardinality of intermediate results). If the data itself is widely distributed, we thus face a problem of having to compute histograms and other statistical synopses over a large number of network nodes each of which holds some data fragment.

In joint work with the University of Patras, Greece, we have developed a framework, suite of distributed algorithms, and software toolkit for approximate computing of distributed statistics with particular emphasis on scalability [4]. This includes support for several kinds of aggregate query processing and the construction, maintenance, and lookup of various histogram types [2]. Our algorithms are layered on top of a new type of *distributed hash sketches* that we developed [3]. This scalable data structure in turn harnesses the basic operations of distributed hash tables (DHT's) such as Pastry or Chord [5, 6, 7]. The DHT infrastructure also provides resilience to network dynamics, the so-called churn problem.

Extensive experiments show our approach does indeed reconcile scalable behavior with low overhead and fairly good accuracy, for a wide range of statistical estimations over widely distributed data.

References

[1] J. M. Hellerstein, P. J. Haas, and H. J. Wang. Online aggregation. In J. Peckham, ed., *SIGMOD Conference*, 1997, pp. 171–182. ACM Press.

[2] Y. E. Ioannidis. The history of histograms (abridged). In *VLDB*, 2003, pp. 19–30.

- [3] N. Ntarmos, P. Triantafillou, and G. Weikum. Distributed hash sketches: Scalable, efficient, and accurate cardinality estimation for distributed multisets. *ACM Transactions on Computer Systems*, 27(1):1–53, 2009.

- [4] N. Ntarmos, P. Triantafillou, and G. Weikum. Statistical structures for Internet-scale data management. *The VLDB Journal*, 18, 2009. Available via Springer online-first: http://www.springerlink.com/content/a40238582233706r/.

[5] A. I. T. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In R. Guerraoui, ed., *Middleware*, 2001, *LNCS 2218*, pp. 329–350. Springer.

[6] I. Stoica, R. Morris, D. R. Karger, M. F. Kaashoek, and H. Balakrishnan. CHORD: A scalable peer-to-peer lookup service for Internet applications. In *SIGCOMM*, 2001, pp. 149–160.

[7] I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan. Chord: a scalable peer-to-peer lookup protocol for Internet applications. *IEEE/ACM Trans. Netw.*, 11(1):17–32, 2003.

31.8.8 Peer-to-Peer Search for Audio-visual Data

Investigators: Mouna Kacimi, Edwin Lewis-Kelham, Tom Crecelius, and Gerhard Weikum

Web search is dominated today by search giants, such as Google and Yahoo, that deploy a centralized approach using text-only indexes. Consequently, while it is possible to search for audio-visual content, this functionality still is limited to associated text and metadata annotations. Additionally, supporting real content-based audio-visual search requires media specific understanding and extremely high CPU utilization which would not scale in today's centralized solutions.

SAPIR is a European project that addresses the problems mentioned above. It aims at developing a large-scale, peer-to-peer (P2P) architecture for searching audio-visual content using the query-by-example paradigm. As our main task in this project, we have designed the system architecture of SAPIR [4]. This entails identifying core components and their key concepts, defining the interfaces and the interplay of components, as well as defining detailed APIs for program development. The architecture addresses three major challenges: scalability, versatility, and extensibility and adaptability. Its components cover the functionalities needed for content analysis and enrichment, index management, caching and replication, P2P statistics, social network management, query routing and processing, result ranking, and UI and device adaptation. We have defined four main components in the SAPIR architecture.

- *Networking*: provides functional means of interaction between peers via one or more overlay networks. It provides communication protocols and network routing functionalities. The SAPIR architecture handles multiple types of overlay networks for different types of media and indexing techniques. Additionally, it supports different types of devices such as PCs and mobile phones. The goal is to have peers with several capabilities to achieve different tasks such as features extraction, crawling and searching.

- *Content Management*: includes content production, analysis, and enrichment. It provides capabilities for analyzing the multimedia content (text, image, audio and video) of peers and extracting features depending on the type of the data. Feature extraction can be based on text descriptions or media specifications including contours and color distributions in images, video segmentation into scenes and the selection of characteristic frames, or speech-to-text extraction [1]

- *P2P Indexing*: provides information about where content can be found in the network of peers. Indexes are distributed among peers so as to balance the access load, provide scalable throughput, and ensure good response time. The granularity of indexes can be based either on peers or on content-object information. If indexes have peer granularity, we talk about *peer-centric overlays* with autonomous peers where every peer can compile its own data without any global control. When indexes have content-object granularity, we refer to *network-centric overlays* in which peers become storage nodes by committing resources to the global network.

- *P2P Search*: provides querying functionality in the SAPIR architecture. The goal is to find high-quality search results with respect to recall and precision. Tow main types of search can be distinguished; search within overlay networks and search across multiple overlay networks. To handle both types of search, we develop query processing and routing strategies to solve queries in a single overlay network. Additionally, we develop a query analyzer that processes queries to multiple overlays, by decomposing each query and sending subqueries to relevant networks, then merging the returned results.

We have defined detailed APIs that reflect the SAPIR architecture components and provide guidance for the interactions (control flow, data flow) among them, without unduly restricting different implementations. These APIs have been implemented, within the SAPIR project, by multiple overlay networks that have different structures and deal with different types of media including text, speech, images, music and video. MINERVA [2, 3] is our main contribution to the SAPIR implementation. It is used as a peer-centric overlay network for indexing audio-visual content based on text descriptions. Minerva combines local index structures of autonomous peers with a global index based on a distributed hash table as an overlay network. The data collection of a peer is locally indexed using inverted lists, where for each key (e.g., keyword, term, or tag) is associated the list of the corresponding URLs of images or other media objects. Minerva maintains a global index that holds very compact, aggregated summaries of the peers' local indexes. The global index implementation is based on Pastry [6], a freely available implementation of a distributed hash table (DHT).

For determining the peers that collaborate on processing a search request, advanced query routing techniques can be used [5]. A query initiator selects a few most promising peers based

on their published per-term summaries, e.g., by executing a distributed top-k algorithm. Subsequently, it forwards the complete query to the selected peers which execute the query locally. This query execution does not involve a distributed top-k query execution since each peer maintains a full-fledged local index with all information necessary to execute the query locally. Finally, the results from the various peers are combined at the querying peer into a single result list. Search results provided by Minerva are merged with results from other overlay networks responsible for multimedia features. The final results are then presented to the user.

References

- [1] W. Allasia, F. Falchi, F. Gallo, M. Kacimi, A. Kaplan, J. Mamou, Y. Mass, and N. Orio. Audio-visual content analysis in P2P networks: The SAPIR approach. In *The 19th International Conference on Database and Expert Systems Application (DEXA 2008)*, Turin, Italy, 2008, pp. 610–614. IEEE Computer Society.

- [2] M. Bender, T. Crecelius, S. Michel, and J. X. Parreira. P2P Web search: Make it light, make it fly (demo). In *3rd Biennial Conference on Innovative Data System Research (CIDR 2007)*, Asilomar, USA, 2007, pp. 164–168. www.crdrdb.org.

[3] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. MINERVA: Collaborative P2P search (demo). In K. Böhm, C. S. Jensen, L. M. Haas, M. L. Kersten, P.-A. Larson, and B. C. Ooi, eds., *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005)*, Trondheim, Norway, 2005, pp. 1263–1266. ACM.

- [4] F. Falchi, M. Kacimi, Y. Mass, F. Rabitti, and P. Zezula. SAPIR: Scalable and distributed image searching. In *SAMT (Posters and Demos)*, Genoa, Italy, 2007, *CEUR Workshop Proceedings*, vol. 300, pp. 11–12. CEUR-WS.org.

[5] S. Michel, M. Bender, N. Ntarmos, P. Triantafillou, G. Weikum, and C. Zimmer. Discovering and exploiting keyword and attribute-value co-occurrences to improve P2P routing indices. In P. S. Yu, V. J. Tsotras, E. A. Fox, and B. Liu, eds., *ACM 15th Conference on Information and Knowledge Management (CIKM2006)*, Arlington, USA, 2006, pp. 172–181. ACM. Acceptance Ratio: 1:6.

[6] A. I. T. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In R. Guerraoui, ed., *Middleware*, 2001, *LNCS 2218*, pp. 329–350. Springer.

31.8.9 Distributed Harvesting and Partitioned Indexing for Web Archiving

Investigators: Avishek Anand, Srikanta Bedathur, Ralf Schenkel, and Christos Tryfonopoulos

The World Wide Web has become a key resource of information related to almost all walks of life, and, at the same time, is rapidly evolving. The resulting ephemeral nature of born-digital content on the Web has necessitated the efforts to archive its contents by organizations like the Internet Archive or the European Archive. These archives are invaluable for historical analyses as they capture the timelines of entities and topics over time, and for reflecting the trends of our society, economy, and culture.

Human-assisted Harvesting of Archives

Despite advances in Web archiving [6], the state-of-the-art is still far away from reaching its ultimate objective of preserving all digital-born information on the Web for posterity. This is due to the lack of large-scale archive gathering to cover the rapidly changing Web but, more importantly, due to limitations imposed by Web servers on crawlers to prevent them from overloading their sites.

However, these limitations are clearly not applicable to normal users of these sites. Based on this observation, we have developed a new notion of archiving called *Human Assisted Harvesting of Archives* (HAHA). Here, we augment the traditional archival crawls with human-assisted captures, by building browser or proxy plugins that capture the history of documents visited by the user. This history information consists of the visit-timestamp, URL, content and content-signatures. Periodically, these local histories are offloaded onto a central archival repository. An additional benefit of this approach is that "hot" and upcoming regions of the Web, which would have been missed by traditional archival crawls, can be captured since Web users can locate such parts.

The effectiveness of such an approach depends on a sufficient number of users visiting a dynamic Web site to capture the site or individual pages at different time points in its evolution. Studies have shown that close to 50% of an individual's browsing activity is page-revisits [5]. Thus, even a single user has a high chance of capturing many versions of a single page. Retaining such a history of Web pages visited by the user for later-time analysis has attracted some attention for personal archiving [1]. Our initial studies [2, 3] were conducted using the publicly available access logs of the 1998 FIFA Soccer Worldcup site [4] with more than 1.3 billion accesses in a span of 88 days. We make the following observations: The current-day archival crawlers capture about 12% of all the page-versions. On the other hand, with less than 5% of Web site visitors participating in the HAHA effort, we can achieve more than 40% version coverage. These results show that our proposed approach can be an effective way to improve the quality and coverage of Web archives.

Partitioned Indexing for Web Archiving

Moreover, we defined a novel framework for range-partitioning the postings of an inverted text index over the various versions of a Web page, using time intervals as criterion for creating the partitions [2]. Our approach naturally entails a certain degree of redundancy between versioned postings as—for efficient time-range queries—each posting needs to be replicated among all partitions that intersect the posting's lifespan. Based on this setup, we defined an interesting optimization problem of how to choose index partitions such that the number of recoverable versions is maximized subject to a bound on the overall space consumption. In a dynamic peer-to-peer environment, recoverability refers to being able to tolerate the loss of an index partition caused by a peer failure. We have shown that this problem is NP-hard in general, and therefore developed several heuristic algorithms that perform very well in experimental studies using the Wikipedia version history as underlying corpus.

References

[1] E. Adar, M. Dontcheva, J. Fogarty, and D. Weld. Zoetrope: Interacting with the ephemeral Web. In *Proceedings of the ACM UIST*, 2008.

- [2] A. Anand. Index partitioning strategies for peer-to-peer Web archival. Masters thesis, Universität des Saarlandes, 2009.
- [3] A. Anand, S. Bedathur, K. Berberich, R. Schenkel, and C. Tryfonopoulos. EverLast: A distributed architecture for preserving the Web. In *Proceedings of the Joint Conference on Digital Libraries (JCDL 2009)*, Austin, Texas, 2009.

[4] M. Arlitt and T. Jin. 1998 World Cup Site Access Logs. http://www.acm.org/sigcomm/ITA/, 1998.

[5] E. Herder. Characterizations of user Web revisit behavior. In *Proceedings of the Workshop on Adaptivity and User Modeling in Interactive Systems*, 2005.

[6] J. Masanès. *Web Archiving*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

31.8.10 Transparent Recovery of Data and Services

Investigators: Gerhard Weikum in cooperation with Alan Fekete, David Lomet, German Shegalov, and Mike Zwilling

To protect mission-critical data from becoming inconsistent because of software, hardware, or environmental failures, mature systems like database and transaction processing systems are built upon fail-stop behavior and have a sophisticated recovery manager. When self-inspection suspects an error, the system forces itself to an immediate shutdown; upon the system's automated restart, recovery procedures are executed to reconstruct the data in its most recent consistent state. For database recovery, the algorithms, their efficiency aspects, and the system engineering are fairly well understood. In distributed settings like web services or cloud computing with interactions of many autonomous components, the additional issues of how to recover messages, session states, and process states, have been much less studied. In particular, general and efficient composability of per-component recovery mechanisms has been an open issue for a long time. Our earlier work on Interaction Contracts [1] has developed a general framework for building comprehensive recovery protocols between components in a systematic, relatively simple, but efficient manner. This work, which was jointly carried out with the database group of Microsoft Research in Redmond, led to software prototypes and also to new results on formal verification of advanced recovery protocols using model-checking techniques. The latter have been published in [4, 5].

More recently, a new direction is being explored based on the Interaction-Contracts framework, motivated by the proliferation of cloud computing. Cloud platforms promise ultra-scalable, zero-administration, and extremely energy- and cost-efficient data-management services, with very easy and flexible deployment of new applications and composable services. It should be easy to add and integrate specialized functions, e.g., for querying and updating RDF data, or specialized access methods, e.g., for 3D shapes or animated avatars, into Web 2.0 applications such as social communities or virtual worlds. Non-functional guarantees like recovery should be automatically provided, ideally transparently to the application programming.

These ambitious goals are in tension with the traditional architectures of database systems where storage and indexing methods are highly integrated with transactional concurrency control and recovery for performance reasons [2]. New data services would either have to implement their own recovery, which is highly time-consuming and error-prone, or could not have transaction-based data-consistency guarantees, or would have to use inefficient protocols with expensive interactions between coarse-grained components.

In joint work involving Microsoft Research, the University of Sydney, and the Max Planck Institute for Informatics, we have designed a new approach [3] to overcome this dilemma. We view data-management services in the cloud as a set of data components (DC's) that interact with transaction components (TC's) using extended forms of the Interaction Contracts developed in [1]. TC's work at a logical level only: they know about transactions and their "logical" concurrency control and undo/redo recovery, but do not know about storage pages, disk caching, B-trees, etc. A DC knows about the physical storage structure of its specific data. It supports record- or object-style access operations which are guaranteed to be atomic, but it does not know about transactions for composing multiple operations into an atomic unit. Providing atomic operations may itself involve DC-local concurrency control and recovery, with more straightforward implementations. The interaction of the mechanisms in TC and DC leads to multi-level redo algorithms, as opposed to the established physical-redo approach of integrated database engines. [3] gives a complete set of protocol requirements – the actual contracts between TC's and DC's – and an outline of appropriate techniques for efficient implementation. A prototype implementation is planned for fall 2009.

References

[1] R. Barga, D. Lomet, G. Shegalov, and G. Weikum. Recovery guarantees for Internet applications. *ACM Transactions on Internet Technology*, 4(3):289–328, 2004.

[2] J. M. Hellerstein, M. Stonebraker, and J. R. Hamilton. Architecture of a database system. *Foundations and Trends in Databases*, 1(2):141–259, 2007.

[3] D. B. Lomet, A. Fekete, G. Weikum, and M. J. Zwilling. Unbundling transaction services in the cloud. In *CIDR*, 2009. www.crdrdb.org.

- [4] G. Shegalov and G. Weikum. Formal verification of a transactional interaction contract. In *IEEE SOA Industry Summit (SOAIS 2008)*, Honolulu, Hawaii, USA, 2008, pp. 525–528. IEEE Computer Society.

- [5] G. Shegalov and G. Weikum. Formal verification of Web service interaction contracts. In *2008 IEEE International Conference on Services Computing (SCC 2008)*, Honolulu, Hawai, USA, 2008, pp. 525–528. IEEE Computer Society.

31.9 Efficient Search in Semistructured Data Spaces (Associated Independent Group)

Coordinator: Ralf Schenkel

31.9.1 Search and Recommendation in Social Tagging Networks

Investigators: Tom Crecelius, Mouna Kacimi, and Ralf Schenkel

Online communities like Flickr, del.icio.us and LibraryThing have established themselves as very popular and powerful services for publishing and searching contents, but also for identifying other users who share similar interests. In these communities, data are usually annotated with carefully selected and often semantically meaningful tags. Items like URLs, photos, videos, or books are typically retrieved by issuing queries that consist of a set of tags, returning items that have been frequently annotated with these tags. However, users often prefer a more personalized way of searching over such a 'global' search, exploiting preferences of and connections between users.

The SENSE system (for *S*ocially *EN*hanced *S*earch and *E*xploration) [2, 3] that we have developed supports personalization along two dimensions: in the social dimension, a search process is focused towards items tagged by users explicitly selected as friends by the querying user, whereas in the spiritual dimension, users that share preferences with the querying user ("brothers in spirit") are preferred [1, 5]. Orthogonal to this, the system additionally integrates semantic expansion of query tags to improve search results. Experimental evaluation with small-scale user studies on data from Flickr and LibraryThing have shown that such a socially aware search often retrieves better results than standard search that solely relies on global frequency of tags [6].

To efficiently evaluate queries in large-scale social networks, SENSE provides a top-k algorithm that dynamically expands the search to related users and tags, extending the well-known Threshold Algorithm [4] for the graph-based aggregation used in SENSE. It relies on two different kinds of precomputed lists of related users: one that maintains related users in the social networks in descending order of distance (for the social dimension), and one that keeps users with similar items in descending order of similarity (for the spiritual dimension). The algorithm reads from these lists and standard per-tag inverted lists and incrementally builds results candidates with score bounds, pruning away candidates that cannot quality for the final results as early as possible and stopping as soon as it can be guaranteed that the results have been found. In an extensive experimental evaluation based on data from Flickr, del.icio.us and LibraryThing, we could demonstrate performance gains of up to a factor of five compared to a join-then-sort algorithm [5], providing interactive response times even for queries with many tags.

References

- [1] M. Bender, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum. Exploiting social relations for query expansion and result ranking. In *Data Engineering for Blogs, Social Media, and Web 2.0, ICDE 2008 Workshops*, Cancun, Mexico, 2008, pp. 501–506. IEEE Computer Society.

- [2] T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum. Making sense: Socially enhanced search and exploration. In *Proceedings of the 34th International Conference on Very Large Data Bases (VLDB 2008)*, Auckland, New Zealand, 2008, vol. 1, pp. 1480–1483. ACM.
- [3] T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum. Social recommenations at work (demo). In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2008)*, Singapore, Singapore, 2008, pp. 884–884. ACM.

[4] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *J. Comput. Syst. Sci.*, 66(4):614–656, 2003.

- [5] R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, and G. Weikum. Efficient top-k querying over social-tagging networks. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2008)*, Singapore, Singapore, 2008, pp. 523–530. ACM.
- [6] R. Schenkel, T. Crecelius, M. Kacimi, T. Neumann, J. Parreira, M. Spaniol, and G. Weikum. Social wisdom for search and recommendation. *IEEE Data Engineering Bulletin*, 31(2):40–49, 2008.

31.9.2 Proximity Search in Text and XML

Investigators: Andreas Broschart, Ralf Schenkel, and Martin Theobald

In search engines, scoring functions play an important role to rank results of user queries. The quality of the scoring function is decisive for user satisfaction and thw success of the search engine. Content-based scoring functions make use of the "bag of words" model, but seem to make little use of *term proximity* information, i.e. the distance between query term occurrences in a document. Suppose a user asks the query "surface area of a triangular pyramid". Scoring functions that completely ignore proximity information may consider documents relevant that contain query terms frequently but in different paragraphs that are likely to treat different topics. For example, in a document related to geometric objects, the first paragraph might elaborate on the "volume of a triangular prism", while the second talks about the "volume of a square pyramid", and the third about the "surface area of a cylinder". Each of the query terms will individually occur quite frequently but not in the user-intended context. From a user's point of view formulating her information need as a phrase query, thus explicitly requesting adjacency of terms, might be a solution to prevent such results. Unfortunately this comes at the expense of many discarded good results – documents carrying information about the "surface area of a pyramid composed of four triangular faces" would certainly be a good hit, but excluded by the phrase query. Proximity scores provide a solution to alleviate those effects by supporting also "soft phrases" without the need to specify exact phrase boundaries by the user. As existing approaches in the field of proximity scores such as [4, 8] focus on retrieval quality but not on efficiently retrieving such results, it is often unclear how to use these scores in a real search engine.

Text Retrieval

In [7], we presented an efficient evaluation framework that relies on our modified version of Büttcher's scoring model [4] integrated into TopX [9] (see Subsection 31.6.1), which aims to

quickly retrieve the top-k results using the family of threshold algorithms (TA) [5].

We have chosen Büttcher's approach as in our earlier studies it was one of the very few proximity-enhanced scoring models which showed significant improvements in result quality compared to Okapi BM25 [6]. Büttcher linearly combines the content score BM25 with a proximity score for each query term into a proximity-aware document-level score. To make index construction feasible, our enhanced version of the proximity score part considers *every* query term occurrence, *not only adjacent* query terms as proposed in [4]. Our extension allows efficient precomputation of indexes without knowing the query load at index construction time while the result quality of queries remains comparable to the original approach.

In [7], we carried out our experiments with the 100 Ad Hoc topics from the 2004 and 2005 TREC Terabyte tracks on the TREC Terabyte collection which consists of about 25 million documents in 426 GB of data. In our experiments we showed that the query processing costs imposed by mere content score index lists can be lowered by one order of magnitude through the use of additional combined index lists. The latter contain for an unordered term pair both content scores and proximity scores and are ordered by proximity score. However, the index size would get prohibitively large if one tried to materialize it. Hence, we pruned the index structures and discovered that already small list lengths (e.g., the best 2,000 postings) can provide a result quality which is comparable to the quality provided by TopX on complete index lists. Pruning not only keeps index sizes manageable but also lowers the query processing costs by an additional order of magnitude. All in all, compared to the unpruned content score lists, the use of pruned content score index lists together with pruned combined index lists accelerates the query execution by two orders of magnitude. At the same time it keeps up the result quality that one gets when using the unpruned content score along with the unpruned combined index lists.

In [1], we exploited these results by proposing a merge-join-based approach that uses pruned index lists. That way we keep up the quality of the retrieved results but get rid of the overhead costs induced by TA. These include maintaining the candidate pool and computing best/worst scores to finally come up with the top-k results. We cut the index lists at a certain threshold and reordered them in document order. The resulting lists can be combined using a merge join operation which is followed by a sort operation by scores. This approach is much easier to implement and needs less memory during evaluation.

XML Retrieval

In [2], we presented a proximity score model geared towards answering content-only queries on XML data. The approach is based on our modified version [7] of Büttcher's scoring model [4]. In order to use this text retrieval scoring model in XML retrieval, it linearizes the text content of an XML document. Moreover our approach takes into account the document structure by optional insertion of gaps at the borders of selected element types to increase distances between text content in unrelated element types. Experiments with 111 assessed content-only topics from the INEX 2006 Ad Hoc Track showed that – like in text retrieval – proximity almost always helps to improve the retrieval quality compared to the purely content based approach with Okapi BM25 scoring. The results for additional gaps have been mixed; learning appropriate gap sizes using relevance feedback techniques is part of our future work. We submitted this approach to INEX 2008 [3] and ranked among the top-3 out of 61 submissions

in the Focused Task of the Ad Hoc Track. The evaluation confirmed our claim from [2] that the result quality is improved by the proximity-based approach.

References

- [1] A. Broschart and R. Schenkel. Effiziente Textsuche mit Positionsinformation. In H. Höpfner and F. Klan, eds., *20. Workshop über Grundlagen von Datenbanken*, Apolda, Germany, 2008, pp. 101–105. International University in Germany.
- [2] A. Broschart and R. Schenkel. Proximity-aware scoring for XML retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2008)*, New York, USA, 2008, pp. 845–846. ACM.
- [3] A. Broschart, R. Schenkel, and M. Theobald. Proximity-aware scoring for XML retrieval. In S. Geva, J. Kamps, and A. Trotman, eds., *Preproceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, Schloss Dagstuhl, Germany, 2008, pp. 46–49.

[4] S. Büttcher, C. L. A. Clarke, and B. Lushman. Term proximity scoring for ad-hoc retrieval on very large text collections. In *Proc. of the 30th Annual Int. ACM SIGIR Conf. on Research & Development in Information Retrieval (SIGIR 2006), Seattle, USA*, 2006, pp. 621–622.

[5] I. F. Ilyas, G. Beskales, and M. A. Soliman. A survey of top-k query processing techniques in relational database systems. *ACM Comput. Surv.*, 40(4), 2008.

[6] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. The Third Text REtrieval Conference (TREC-3). In *TREC*, 1994, pp. 109–126.

- [7] R. Schenkel, A. Broschart, S. Hwang, M. Theobald, and G. Weikum. Efficient text proximity search. In N. Ziviani and R. A. Baeza-Yates, eds., *14th String Processing and Information Retrieval Symposium (SPIRE 2007)*, Santiago, Chile, 2007, *LNCS 4726*, pp. 287–299. Springer.

[8] R. Song, M. J. Taylor, J.-R. Wen, H.-W. Hon, and Y. Yu. Viewing term proximity from a different perspective. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, eds., *ECIR*, 2008, *LNCS 4956*, pp. 346–357. Springer.

- [9] M. Theobald, H. Bast, D. Majumdar, R. Schenkel, and G. Weikum. TopX: Efficient and versatile top-k query processing for semistructured data. *The VLDB Journal*, 17(2):81–115, 2008.

31.9.3 Top-k Queries with Resource Constraints

Investigators: Ralf Schenkel and Gerhard Weikum in cooperation with Michal Shmueli and Yosi Mass

An important class of queries frequently used in many applications areas are top-k queries that seek, possibly from a large set of data items, those k items that have the highest aggregated values in certain dimensions. Important examples for such application areas are text retrieval based on per-term score values, image retrieval based on features with scores, retrieving the most important events from a stream of sensor readings data, or finding network clients generating the highest load from a set of network statistics. Simple solutions that first aggregate the corresponding values for each item and then sort all items to find those with highest aggregations are way too expensive for huge collections, so the top-k problem has been widely studied in the literature. The current state-of-the-art solution for

this problem is the family of threshold algorithms (TA) proposed by Fagin et al. [2] that utilizes precomputed lists that contain, for each query dimension, items in descending order of their value in that dimension. One of the most efficient algorithms within this framework is our IO-Top-k algorithm [1]. It combines a block-based structure of the precomputed lists with benefit-based scheduling of sequential accesses to these blocks and a cost-based heuristic for scheduling random accesses to fully evaluate the score of promising result candidates that have been seen only in a few lists before.

However, in some cases only limited resources can be utilized to process a query. For example, a user of a search engine typically wants to wait only a few seconds for a result of a query and will cancel her search operation when it takes longer. Other limits may be imposed by saving battery power in sensors by allowing only a certain maximal number of operations per query, or by bandwidth constraints of a network. Existing algorithms for top-k processing do not take such a resource budget into account when making scheduling decisions, but focus on the performance to compute the final top-k results. Under budget constraints, these *budget-oblivious* algorithms often return results that are a lot worse than the results that can be achieved with a clever, budget-aware scheduling.

We have developed a set of novel algorithms for best-effort top-k query processing given a fixed budget limit for the execution cost, where the budget can be defined in terms of time, number of disk accesses, or number of network messages, reflecting an essential limitation faced by contemporary applications [3]. Our *budget-aware* top-k processing produces results that are significantly better than those of state-of-the-art budget-oblivious solutions.

References

[1] H. Bast, D. Majumdar, R. Schenkel, M. Theobald, and G. Weikum. IO-Top-k: Index-access optimized top-k query processing. In U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim, eds., *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB 2006)*, Seoul, Korea, 2006, pp. 475–486. ACM. Acceptance ratio 1:7.

[2] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *J. Comput. Syst. Sci.*, 66(4):614–656, 2003.

- [3] M. Shmueli-Scheuer, C. Li, Y. Mass, H. Roitman, R. Schenkel, and G. Weikum. Best effort top-k query processing under budgetary constraints. In Y. Ioannidis, D. Lee, and R. Ng, eds., *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE 2009)*, Shanghai, China, 2009.

31.9.4 Relevance Feedback in XML Search and Exploration

Investigators: Ralf Schenkel and Hanglin Pan

Relevance Feedback is an important way to enhance retrieval quality by integrating relevance information provided by a user. In XML retrieval, most existing feedback engines usually generate an expanded keyword query from the content of elements marked as relevant or non-relevant. This approach, which is inspired by text-based IR, completely ignores the semistructured nature of XML.

We have been among the first to propose relevance feedback methods for XML retrieval beyond simple keyword-based query expansion. Out methods exploit the structure of results

to generate expanded structural queries, which yield more precise results than the original query [3]. Our more recent work [1] considers advanced user feedback beyond simple per-result feedback. Instead of supplying a single relevance value for a complete result element, the user may give feedback with different granularities and types. For example, the user may like a section in an article for content, but dislike the entire article for structure and also dislike a specific paragraph for content within the good section. We maintain a pool of possible query refinements (i.e., reweighing, adding or removing terms, structural constraints, or ontological expansions). Following earlier work by Ruthven and Lalmas [2] for text retrieval, we combine for each possible refinement candidate the relevance value for the result as delivered by the search engine, the initial weight of the candidate and the user feedback (with tunable weights according to granularity and type) using the Dempster-Shafer theory of evidence [4]. We then use the Transferable Belief Model [5] to compute, for each refinement candidate, a probability that a query refinement can identify relevant results.

References

- [1] H. Pan, R. Schenkel, and G. Weikum. Fine-grained relevance feedback for XML retrieval (demo). In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2008)*, Singapore, Singapore, 2008, pp. 887–887. ACM.

[2] I. Ruthven and M. Lalmas. Using dempster-shafer's theory of evidence to combine aspects of information use. *Journal of Intelligent Information Systems*, 19(3):267–302, 2002.

[3] R. Schenkel and M. Theobald. Structural feedback for keyword-based XML retrieval. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, eds., *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006*, London, UK, 2006, *LNCS 3936*, pp. 326–337. Springer.

[4] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.

[5] P. Shets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–243, 1994.

31.10 Academic Activities

31.10.1 Journal Positions

Gerhard Weikum:

- *Member of the Editorial Board of Foundations and Trends in Databases* (since 2009)
- *Member of the Editorial Board of Communications of the ACM* (since 2008)
- *Member of the Editorial Board of IEEE Transactions on Knowledge and Data Engineering* (2003–2007)

31.10.2 Book Positions

Gerhard Weikum:

- *Member of the Editorial Board of Springer Lecture Notes in Computer Science (LNCS)* (since 2004)

31.10.3 Conference and Workshop Positions

Membership in Program Committees

Ralitsa Angelova:

- *3rd International ICST Conference on Scalable Information Systems (INFOSCALE)*, Vico Equense, Italy, June 2008
- *23rd ACM Symposium on Applied Computing (SAC)*, Engineering Large-Scale Distributed Systems Track, Fortaleza, Brazil, March 2008
- *2nd International ICST Conference on Scalable Information Systems (INFOSCALE)*, Suzhou, China, June 2007

Srikanta Bedathur:

- *18th International World Wide Web Conference (WWW)*, Madrid, Spain, April 2009
- *35th International Conference on Very Large Data Bases (VLDB)* (demonstrations track & developers track), Lyon, France, August 2009
- *14th International Conference on Management of Data (COMAD)*, Bombay, India, December 2008

Andreas Broschart:

- *Initiative for the Evaluation of XML Retrieval (INEX)*, Schloß Dagstuhl, December 2008

Thomas Neumann:

- *25th IEEE International Conference on Data Engineering (ICDE)*, Hongkong, April 2009
- *24th IEEE International Conference on Data Engineering (ICDE)*, Cancún, México, April 2008
- *13th Conference Datenbanksysteme für Business, Technologie und Web (BTW)*, Münster, Germany, March 2009
- *11th International Conference on Extending Database Technology (EDBT)*, Nantes, France, March 2008
- *17th ACM Conference on Information and Knowledge Management (CIKM)*, Napa Valley, USA, October 2008
- *11th East-European Conference on Advances in Databases and Information Systems (ADBIS)*, Varna, Bulgaria, September 2007

Josiane Xavier Parreira:

- *6th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR)*, Napa Valley, USA, October 2008
- *23rd ACM Symposium on Applied Computing (SAC)*, Engineering Large-Scale Distributed Systems Track, Fortaleza, Brazil, March 2008

Maya Ramanath:

- *18th International World Wide Web Conference (WWW)*, Madrid, Spain, April 2009
- *17th ACM Conference on Information and Knowledge Management (CIKM)*, Napa Valley, USA, October 2008
- *17th International World Wide Web Conference (WWW)*, Beijing, China, April 2008
- *16th International World Wide Web Conference (WWW)*, Banff, Alberta, Canada, May 2007

Ralf Schenkel:

- *26th IEEE International Conference on Data Engineering (ICDE)*, Long Beach, USA, March 2010
- *17th International Conference on Cooperative Information Systems (CoopIS)*, Vilamoura, Portugal, November 2009
- *Workshop of the GI Special Interest Group Information Retrieval*, Darmstadt, Germany, September 2009
- *6th International Workshop on Text-Based Information Retrieval (TIR)*, Linz, Austria, August 2009
- *ACM SIGMOD International Conference on Management of Data*, Demo Track, Rhode Island, USA, June 2009
- *25th IEEE International Conference on Data Engineering (ICDE)*, Shanghai, China, April 2009
- *31st European Conference on Information Retrieval (ECIR)*, Toulouse, France, March 2009
- *12th International Conference on Extending Database Technology (EDBT)*, Demo Track, St. Petersburg, Russia, March 2009
- *13th Conference Datenbanksysteme für Business, Technologie und Web (BTW)*, Münster, Germany, March 2009
- *Workshop Database as a Service*, Münster, Germany, March 2009
- *4th International Workshop on Database Technologies for Handling XML Information on the Web (DataX)* St. Petersburg, Russia, March 2009
- *Initiative for the Evaluation of XML Retrieval (INEX)*, Schloß Dagstuhl, December 2008
- *16th International Conference on Cooperative Information Systems (CoopIS)*, Monterrey, USA, November 2008
- *34th International Conference on Very Large Databases (VLDB)*, Auckland, New Zealand, September 2008
- *19th International Conference on Database and Expert Systems Applications (DEXA)*, Turin, Italy, September 2008
- *Workshop of the GI Special Interest Group Information Retrieval*, Würzburg, Germany, September 2008

- *31st International Conference on Research and Development in Information Retrieval (SIGIR)*, Singapore, July 2008
- *30th European Conference on Information Retrieval (ECIR)*, Glasgow, U.K., March 2008
- *11th International Conference on Extending Database Technology (EDBT)*, Nantes, France, March 2008
- *3rd International Workshop on Database Technologies for Handling XML Information on the Web (DataX)* Nantes, France, March 2008
- *Initiative for the Evaluation of XML Retrieval (INEX)*, Schloß Dagstuhl, December 2007
- *15th International Conference on Cooperative Information Systems (CoopIS)*, Vilamoura, Portugal, November 2007
- *33rd International Conference on Very Large Databases (VLDB)*, Demo Track, Vienna, Austria, September 2007
- *18th International Conference on Database and Expert Systems Applications (DEXA)*, Regensburg, Germany, September 2007
- *Workshop of the GI Special Interest Group Information Retrieval*, Halle, Germany, September 2007

Marc Spaniol:

- *9th International Conference on Knowledge Management and Knowledge Technologies (I-Know'09)*, Graz, Austria, September 2009
- *6th International Workshop on Text-based Information Retrieval (TIR 2009)*, Linz, Austria, August 2009
- *8th International Conference on Web-based Learning (ICWL 2009)*, Aachen, Germany, August 2009
- *9th Workshop on Multimedia Metadata (WMM'09)*, Toulouse, France, March 2009
- *Workshop on Social Information Retrieval for Technology-Enhanced Learning (SIRTEL 2008)*, Maastricht, The Netherlands, September 2008
- *Workshop on Story-Telling for Educational Gaming (STEG 2008)*, Maastricht, The Netherlands, September 2008
- *7th International Conference on Web-based Learning (ICWL 2008)*, Jinhua, China, August 2008
- *8th MPEG-7/21 Community Workshop on Multimedia Metadata Management & Retrieval (M3R 08)*, Klagenfurt, Austria, May 2008
- *International Conference on Technology, Communication and Education i-TCE 2008*, Kuwait, April 2008

Martin Theobald:

- *26th IEEE International Conference on Data Engineering (ICDE)*, Long Beach, USA, March 2010

- *3rd Workshop on Personalized Access, Profile Management, and Context Awareness: Databases (PersDB)*, Lyon, France, August 2009
- *32nd ACM SIGIR International Conference on Research and Development in Information Retrieval*, Boston, USA, July 2009
- *25th IEEE International Conference on Data Engineering (ICDE)*, Shanghai, China, April 2009
- *18th International World Wide Web Conference (WWW)*, Madrid, Spain, April 2009
- *4th International Workshop on Database Technologies for Handling XML Information on the Web (DATAX)*, St. Petersburg, Russia, March 2009
- *12th International Conference on Extending Database Technology (EDBT)*, St. Petersburg, Russia, March 2009
- *31st European Conference on Information Retrieval (ECIR)*, Toulouse, France, March 2009
- *Initiative for the Evaluation of XML Retrieval (INEX)*, Schloß Dagstuhl, December 2008
- *2nd Workshop on Personalized Access, Profile Management, and Context Awareness: Databases (PersDB)*, Auckland, New Zealand, August 2008
- *30th European Conference on Information Retrieval (ECIR)*, Glasgow, U.K., March 2008
- *2nd Workshop on Management of Uncertain Data (MUD)*, Auckland, New Zealand, September 2008
- *12th East-European Conference on Advances in Databases and Information Systems (ADBIS)*, Pori, Finland, September 2008
- *2nd International Workshop on Ranking in Databases (DBRank)*, Cancún, México, April 2008
- *3rd International Workshop on Database Technologies for Handling XML Information on the Web (DATAX)*, Nantes, France, March 2008
- *11th International Conference on Extending Database Technology (EDBT)*, Nantes, France, March 2008
- *Initiative for the Evaluation of XML Retrieval (INEX)*, Schloß Dagstuhl, December 2007
- *9th ACM International Workshop on Web Information and Data Management (WIDM)*, Lisboa, Portugal, November 2007
- *1st Workshop on Management of Uncertain Data (MUD)*, Vienna, Austria, September 2007
- *11th East-European Conference on Advances in Databases and Information Systems (ADBIS)*, Varna, Bulgaria, September 2007
- *10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL)*, Corfu, Greece, June 2007
- *29th European Conference on Information Retrieval (ECIR)*, Rome, Italy, April 2007

Christos Tryfonopoulos:

- *1st International Conference on Advances in P2P Systems (AP2PS)*, Sliema, Malta, October 2009
- *20th International Conference on Database and Expert Systems Applications (DEXA)*, Linz, Austria, August 2009
- *6th European Semantic Web Conference (ESWC)*, Heraklion, Greece, May 2009
- *1st International Workshop on Efficiency Issues in Information Retrieval (EEIR)*, Glasgow, Scotland, March 2008
- *7th Hellenic Data Management Symposium (HDMS)*, Heraklion, Greece, July 2008
- *6th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR)*, Napa Valley, USA, October 2008
- *5th European Semantic Web Conference (ESWC)*, Tenerife, Spain, June 2008
- *1st International Workshop on Data Management in P2P Systems (DaMap)*, Nantes, France, March 2008
- *20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Dayton, USA, November 2007
- *11th East-European Conference on Advances in Databases and Information Systems (ADBIS)*, Varna, Bulgaria, September 2007
- *2nd International Conference on Scalable Information Systems (INFOSCALE)*, Suzhou, China, June 2007
- *6th International Semantic Web Conference (ISWC)*, Busan, Korea, November 2007
- *2nd Asian Semantic Web Conference (ASWC)*, Busan, Korea, November 2007
- *5th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR)*, Amsterdam, The Netherlands, July 2007
- *10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL)*, Corfu, Greece, June 2007

Gerhard Weikum:

- *26th IEEE International Conference on Data Engineering (ICDE)* (Program Committee Co-Chair, together with S. Ghandeharizadeh and J. Haritsa), Long Beach, USA, March 2010
- *13th East-European Conference on Advances in Databases and Information Systems (ADBIS)*, Riga, Latvia, September 2009
- *35th International Conference on Very Large Data Bases (VLDB)*, Lyon, France, August 2009
- *32nd ACM SIGIR International Conference on Research and Development in Information Retrieval*, Boston, USA, July 2009
- *25th IEEE International Conference on Data Engineering (ICDE)*, Shanghai, China, April 2009
- *18th International World Wide Web Conference (WWW)*, Madrid, Spain, April 2009

- *17th ACM International Conference on Information and Knowledge Management (CIKM)*, Napa Valley, USA, November 2008
- *34th International Conference on Very Large Databases (VLDB)*, Auckland, New Zealand, September 2008
- *31st ACM SIGIR International Conference on Research and Development in Information Retrieval*, Singapore, July 2008
- *24th IEEE International Conference on Data Engineering (ICDE)*, Cancún, México, April 2008
- *1st ACM International Conference on Web Search and Data Mining (WSDM)*, Stanford, USA, February 2008
- *30th ACM SIGIR International Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, July 2007
- *26th ACM SIGMOD International Conference on Management of Data*, Beijing, China, June 2007
- *3rd Biennial Conference on Innovative Data Systems Research (CIDR)* (Program Committee Chair), Asilomar, USA, January 2007

Membership in Organizing Committees

Srikanta Bedathur:

- *1st International Workshop on Modeling, Managing and Mining of Evolving Social Networks (M3SN)*, Shanghai, China, March 2009 (co-chair & co-organizer)

Ralf Schenkel & Martin Theobald:

- *Efficiency Track at the 2009 Initiative for the Evaluation of XML Retrieval (INEX)*, Brisbane, Australia, December 2009 (co-organizers)
- *Efficiency Track at the 2008 Initiative for the Evaluation of XML Retrieval (INEX)*, Schloß Dagstuhl, Germany, December 2008 (co-organizers)

Marc Spaniol:

- *Workshop on Story-Telling for Educational Gaming (STEG 2008)*, Maastricht, The Netherlands, September 2008 (co-organizer)
- *8th International Conference on Web-based Learning (ICWL 2009)*, Aachen, Germany, August 2009 (co-organizer and PC chair)

Gerhard Weikum:

- *General Co-Chair of the 4th Biennial Conference on Innovative Data Systems Research (CIDR)*, Asilomar, USA, January 2009
- *Co-Organizer of the Dagstuhl Seminar on "Ranked XML Querying"* (with S. Amer-Yahia and D. Srivastava), Schloß Dagstuhl, March 2008

31.10.4 Invited Talks and Tutorials

Klaus Berberich:

- *A Time Machine for Text Search*, Invited Talk, Athens University of Economics and Business (AUEB), Greece, May 2008

Gjergji Kasneci:

- *YAGO-NAGA: Harvesting and Querying Knowledge*, Invited Talk, Microsoft Research Cambridge, U.K., November 2008
- *The Complexity of Reasoning about Pattern-based XML Schemas*, Invited Talk, Technische Universität Dortmund, Germany, June 2007

Thomas Neumann:

- *Skalierbare Verwaltung von Semantic-Web-Daten: Effiziente Indexierung und Anfrageoptimierung*, Invited Talk, Technische Universität München, Germany, April 2009
- *Join Order Optimization in Database Systems*, Invited Talk, Universität Zürich, Switzerland, April 2009
- *Skalierbare Verwaltung von Semantic-Web-Daten: Effiziente Indexierung und Anfrageoptimierung*, Invited Talk, Freie Universität Berlin, Germany, September 2008

Ralf Schenkel:

- *Managing Social Wisdom: Efficient and Effective Search in Social Tagging Networks*, Invited Talk, Dagstuhl Seminar "Interactive Information Retrieval", Schloß Dagstuhl, Germany, March 2009
- *Managing Social Wisdom: Efficient and effective search in social tagging networks*, Invited Talk, Technische Universität München, Germany, March 2009
- *Efficient Ranked Retrieval in Large Collections*, Invited Talk, Universität Zürich, Switzerland, December 2008
- *Managing Social Wisdom: Efficient and Effective Search in Social Tagging Networks*, Invited Talk, kickoff meeting of the cluster of excellence "Multimodal Computing and Interaction", Saarbrücken, Germany, November 2008
- *Efficient Search in Social Networks* (in German), Invited Talk, Meeting of the MPG's BAR, Saarbrücken, Germany, October 2008
- *Efficiency Issues in Information Retrieval*, Tutorial, Herbstschule der GI-Fachgruppe "Information Retrieval", Schloß Dagstuhl, Germany, September 2008
- *"IP" is not always "Internet Protocol"*, Invited Talk, Dagstuhl Perspectives Workshop "Virtual games, interactive hosted services and user-generated content in Web 2.0", Schloß Dagstuhl, Germany, September 2008
- *Relevance Feedback in the TopX Search Engine*, Invited Talk, Dagstuhl Seminar "Ranked XML Querying", Germany, April 2008

Mauro Sozio:

- *SOFIE: A Self-Organizing Framework for Information Extraction*, Google Lab, Zürich, Switzerland, April 2009
- *Near-Optimal Dynamic Replication in Unstructured Peer-to-Peer Networks*, Seminar on Algorithmics, La Sapienza University, Rome, Italy, March 2009
- *Replicas in Unstructured P2P Networks: How Many? Where? - A 30-year perspective on replication*, ETH Winterschool, Monte Verita, Switzerland, November 2007

Marc Spaniol:

- *Web Archiving*, Tutorial, Stuttgart, Germany, March 2009
- *Web Archiving*, Tutorial, Aachen, Germany, November 2008
- *Workshop on Knowledge Visualization and Discovery 2008*, Invited Talk, Graz, Austria, September 2008
- *8th International Web Archiving Workshop (IWAW)*, Invited Talk, Århus, Denmark, September 2008

Fabian Suchanek:

- *YAGO meets SOFIE*, Invited Talk, Microsoft Research Silicon Valley, USA, November 2008
- *YAGO meets SOFIE*, Invited Talk, Institute of Applied Informatics and Formal Description Methods, Universität Karlsruhe, Germany, November 2008
- *LEILA and YAGO*, Invited Talk, Institute of Cognitive Science, Universität Osnabrück, Germany, July 2008
- *YAGO - A Core of Semantic Knowledge*, Invited Talk, L3S Research Center, Hannover, Germany, May 2008
- *YAGO - A Core of Semantic Knowledge*, Invited Talk, German Research Center for Artificial Intelligence, Kaiserslautern, Germany, February 2008

Martin Theobald:

- *SpotSigs - Robust and Efficient Near-Duplicate Detection in Large Web Collections*, Senior Researcher Series Talk, Max-Planck-Institut für Informatik, Saarbrücken, Germany, April 2009
- *Trio - A System for Data, Uncertainty, and Lineage*, Demonstration, Dagstuhl Seminar "Uncertainty Management in Information Systems", Schloß Dagstuhl, Germany, October 2008
- *TopX 2.0 - A (Very) Fast Object-Store for Top-k XPath Query Processing*, Invited Talk, Dagstuhl Seminar "Ranked XML Querying", Schloß Dagstuhl, Germany, March 2008
- *Trio - A System for Data, Uncertainty, and Lineage*, Invited Talk, Database Seminar, L'Università della Svizzera Italiana, Lugano, Italy, March 2008

- *Trio - A System for Data, Uncertainty, and Lineage*, Invited Talk, UC Berkeley Database Group Seminar, USA, October 2007
- *TopX - Efficient and Versatile Top-k Query Processing for Text, Semistructured, and Structured Data*, Invited Talk, Solar and Astrophysics Laboratory Seminar, Lockheed Martin, Palo Alto, USA, August 2007

Christos Tryfonopoulos:

- *Alerting Services in Mobile Environments*, Search Goes Mobile (SGM) Workshop, Bertinoro, Italy, April 2008
- *Information Retrieval and Filtering in Structured Overlay Networks*, University of Peloponnese, Tripolis, Greece, July 2007
- *Architectural Alternatives for Information Filtering in Structured Overlays*, University of Ioannina, Ioannina, Greece, May 2007

Gerhard Weikum:

- *Harvesting, Searching, and Ranking Knowledge from the Web*, Keynote at 2nd ACM International Conference on Web Search and Data Mining (WSDM), Barcelona, Spain, February 2009
- *Harvesting, Searching, and Ranking Knowledge from the Web*, Keynote at Dutch-Belgian Workshop on Information Retrieval, Enschede, The Netherlands, February 2009
- *Harvesting, Searching, and Ranking Knowledge from the Web*, Keynote at Swiss DBTA Workshop on Information Retrieval, Basel, Switzerland, November 2008
- *Harvesting, Searching, and Ranking Knowledge from the Web*, Colloquium Lecture at Microsoft Research, Redmond, USA, June 2008
- *Harvesting, Searching, and Ranking Knowledge from the Web*, Colloquium Lecture at HP Labs, Palo Alto, USA, June 2008
- *Harvesting and Organizing Knowledge from the Web*, Keynote at East European Conference on Advances in Databases and Information Systems (ADBIS), Varna, Bulgaria, October 2007
- *Web, Semantic, and Social Information Retrieval*, Tutorial at EDBT Summer School, Bolzano, Italy, September 2007
- *DB & IR: Both Sides Now*, Keynote at ACM SIGMOD International Conference on Management of Data, Beijing, China, June 2007
- *Harvesting and Organizing Knowledge from the Web*, Keynote at German E-Science Conference (GES), Baden-Baden, Germany, May 20
- *What's Cool and What's Uncool about TA?*, International Workshop on Ranking in Databases, in conjunction with ICDE 2007, Istanbul, Ankara, April 2007

31.10.5 Other Academic Activities

Marc Spaniol:

- Steering Committee Member of the Multimedia Metadata Community

- Member of the International Advisory Board of the Spanish ACM SIG Chapter on Computer Science Education (SIGCSE)

Maya Ramanath:

- Technical Advisory Board Member, MeshLabs Inc., Bangalore, India

Gerhard Weikum:

- *Member of the Coordination Committee of the DFG Priority Program on Scalable Visual Analytics* (since 2008)
- *Elected Reviewer of the German Science Foundation (DFG)*, Fachkollegium 409 (since 2008)
- *Member of the Advisory Board of the European Archive* (since 2008)
- *Member of the Scientific Advisory Board of the Centrum Wiskunde & Informatica (CWI)* (since 2008)
- *ACM SIGMOD Advisory Board* (since 2005)
- *ACM SIGMOD Awards Committee* (2005–2009, Chair in 2008)
- *Member of the Scientific Directorate of Schloß Dagstuhl International Conference and Research Center for Computer Science* (since 2004)
- *Spokesperson of the International Max Planck Research School for Computer Science (IMPRS-CS)* (since 2004)
- *President of the VLDB Endowment* (since 2004)
- *Member of the Scientific Board of the Swiss NCCR on Mobile Information and Communication Systems* (since 2003)
- *Member of the GI DBIS (German Computer Society, Special Interest Group on Database and Information Systems) Dissertation Award Committee* (2008)

31.11 Teaching Activities

Summer Semester 2007

Seminars:

Neumann, Schenkel: Query Processing, Universität des Saarlandes

Winter Semester 2007/2008

Courses:

Neumann: Query Optimization, Universität des Saarlandes
Weikum: Information Retrieval and Data Mining, Universität des Saarlandes

Seminars:

Bedathur, Schenkel: Future of Web Search, Universität des Saarlandes

Summer Semester 2008

Courses:

Neumann: Datenbanksysteme, Universität Heidelberg
Neumann, Schenkel: Informationssysteme, Universität des Saarlandes
Theobald: CS245 Principles of Database Systems, Stanford University

Seminars:

Neumann: Anfrageverarbeitung und Optimierung, Universität Heidelberg

Winter Semester 2008/2009

Seminars:

Neumann, Schenkel: Techniques for Nontraditional Data Management

Bachelor Theses

Completed:

- *Tim Benke:* "Semantic Overlay Networks for Peer-to-peer Web Search", April 2007
- *Daniel Dahrendorf:* "Analysis of Different Replication and Caching Strategies in the P2P Search Engine Minerva", March 2008
- *Johannes Heinz:* "Implementation of an Approximate Information Filtering Approach (MAPS)", February 2008
- *Marco Hüster:* "Formale Beschreibung von Fertigungsprozessen mit der Prozessalgebra CC", September 2007
- *Faraz Makari Manshadi:* "Fast Distributed Algorithm for Replication in Unstructured P2P Networks", October 2008
- *Joachim Müller:* "Komplexe Filterpraedikate in RDF-3X", January 2009
- *Marco Stadtmüller:* "An XPath 2.0 Full-Text Query Parser for the TopX Search Engine", July 2007

Diploma and Master's Theses

Completed:

- *Irem Arikan*: "Exploiting Temporal References in Text Retrieval", March 2009
- *Avishek Anand*: "Index Partitioning Strategies for Peer-to-Peer Web Archival", January 2009
- *Laura Maria Andreescu*: "Pricing Information Goods in an Agent-based Information Filtering System", December 2008
- *Caroline Gherbaoui*: "Similarity Measures for Query Expansion in TopX", October 2008
- *Minko Dudev*: "Personalization of Search on Structured Data", September 2008

- *Stefan Holder*: "Replication in Unstructured Peer-to-Peer Networks with Availability Constraints", September 2008
- *Alexander Prohaska*: "Building Web Query Subsumption Hierarchies", September 2008
- *Richard Socher*: "A Learning-Based Hierarchical Model for Vessel Segmentation", July 2008
- *Levan Kasradze*: "Implementation of a File-based Indexing Framework for the TopX Search Engine", May 2008
- *Christian Langner*: "Text-based Approaches for Content-Based Image Retrieval in a P2P Network", April 2008
- *Ciprian Raileanu*: "Entity Extraction and Contextual Rule Selection in the Medical Domain", February 2008
- *Osama Sammodi*: "Incremental Relevance Feedback for TopX", February 2008
- *Mohammed AbuJarour*: "Efficient XML Query Processing and Full-Text Search", January 2008
- *Karin Heß*: "Automatische Verknüpfung historischer und zeitgenössischer Wörterbücher", September 2007
- *Shady Elbassuoni*: "Adaptive Personalization of Web Search", July 2007
- *Alina Kopp*: "Design and Implementation of an Automatic Semantic Annotation Service", July 2007
- *Gaurav Pandey*: "Relevance Feedback using Query Logs", June 2007
- *Silvana Solomon*: "Evaluation of Relevance Feedback Algorithms for XML Retrieval", June 2007

31.12 Dissertations, Habilitations, Offers, Awards

31.12.1 Dissertations

Completed:

- *Georgiana Ifrim:* "A Statistical Learning Approach to Concept-Based Document Classification", February 2009
- *Julia Luxenburger:* "Modeling and Exploiting User Search Behavior for Information Retrieval", December 2008
- *Fabian Suchanek:* "Automated Construction and Growth of a Large Ontology", December 2008
- *Christian Zimmer:* "Approximate Information Filtering in Structured Peer-to-Peer Networks", October 2008
- *Matthias Bender:* "Advanced Methods for Query Routing in Peer-to-Peer Information Retrieval", July 2007
- *Sebastian Michel:* "Top-k Aggregation Queries in Large-Scale Distributed Systems", July 2007

In preparation:

- *Avishek Anand:* "Efficient Content Capturing and Indexing for Search-Enabled Web Archives"
- *Ralitsa Angelova:* "Graph-based Classification and Clustering of Documents and Entities"
- *Klaus Berberich:* "Efficient Time-Travel Search over Web Archives"
- *Andreas Broschart:* "Towards Top-k-Aware Graph Retrieval"
- *Tom Crecelius:* "Social Peer-To-Peer Web Search"
- *Gerard de Melo:* "Automatic Construction of Large-Scale Lexical Knowledge Bases"
- *Dimitar Denev:* "Models and Methods for Temporal Web Mining"
- *Laura Dietz:* "Probabilistic Models for Unsupervised Ranking in Three Problem Domains"
- *Shady Elbassuoni:* "Ranking in Entity-Relationship Graphs"
- *Gjergji Kasneci:* "Searching and Ranking with Entities and Relationships"
- *Ndapandula Nakashole:* "Parallel Information Extraction"
- *Hanglin Pan:* "Feedback-Driven Query Refinement in Ranked XML Retrieval"
- *Lizhen Qu:* "Fact and Opinion Co-Evolution"
- *Bilyana Taneva:* "Statistical Learning Methods for Preference Analysis"
- *Yafang Wang:* "Temporal Fact Extraction"
- *Josiane Xavier Parreira:* "Decentralized Link Analysis in Peer-to-Peer Web Search Networks"

31.12.2 Awards and Honors

- *Irem Arikan, Srikanta Bedathur, and Klaus Berberich:* Best Late Breaking Result Award at WISDM 2009 (ACM International Conference on Web Search and Data Mining) for their paper "Time Will Tell: Leveraging Temporal Expressions in IR", Barcelona, Spain, February 2009
- *Sebastian Michel:* GI/DBIS (German Computer Society, Special Interest Group on Databases and Information Systems) Dissertation Award 2007/2008, Münster, Germany, March 2009
- *Sebastian Michel:* Otto Hahn Medal of the Max Planck Society 2007, Dresden, Germany, June 2008
- *Marc Spaniol:* Friedrich-Wilhelm-Preis Dissertation Award (RWTH Aachen University), Aachen, Germany, 2008
- *Martin Theobald:* ACM SIGMOD Doctoral Dissertation Award Honorable Mention 2006, Beijing, China, June 2007
- *Martin Theobald:* Otto Hahn Medal of the Max Planck Society 2006, Kiel, Germany, June 2007

- *Martin Theobald:* GI/DBIS (German Computer Society, Special Interest Group on Databases and Information Systems) Dissertation Award 2005/2006, Aachen, Germany, March 2007
- *Gerhard Weikum:* elected into the German Academy of Science and Engineering (acatech), 2008
- *Thomas Neumann and Gerhard Weikum:* invited for the Special "Best of VLDB'08" Issue of The VLDB Journal for their VLDB 2008 paper "RDF-3X: a RISC-style Engine for RDF"
- *Gerard de Melo and Gerhard Weikum:* Best Paper Award at ICGL 2008 (International Conference on Global Interoperability for Language Resources) for their paper "A Machine Learning Approach to Building Aligned Wordnets", Hong Kong, China, January 2008

31.13 Grants and Cooperations

31.13.1 Projects Funded by the European Union (EU)

LiWA

LiWA (Living Web Archives, `http://liwa-project.eu/`) is a STREP (Specific Targeted Research Project) that started in February 2008 and involves 8 partners from academia, industry, and library services (L3S Hannover, European Archive, Hungarian Academy of Sciences, Hanzo Archives Ltd., Instituut voor Beeld en Geluid, National Library of the Czech Republic, Moravian Library, Max Planck Institute for Informatics). The goal of LiWA is the development of next-generation Web archiving technologies. The department 5 of the Max Planck Institute for Informatics coordinates the work on temporal coherence, and contributes to tasks on spam detection, semantic evolution, and system integration. Within LiWA we extend our ongoing foundational research in the area of Web archiving and Web mining. Main research issues are improving the capturing process of Web sites for high-quality archives and the interpretability of contents for later retrieval and analysis. Our novel models and strategies are integrated into the European Archive's Web crawler Heritrix (`crawler.archive.org`).

LivingKnowledge

LivingKnowledge is an Integrated Project (IP) in the broad area of Web mining, investigating the diversity and the evolution of facts, opinions, and bias as expressed in digital media (`http://livingknowledge-project.eu/`). The project started in February 2009 and involves 10 partners from academia, industry, and public archival institutions (University of Trento, Yahoo! Research Barcelona, SORA Institute for Social Research and Analysis, Italian National Inter-University Consortium for Telecommunications, European Archive Foundation, University of Pavia, University of Southampton, Indian Statistical Institute, L3S Research Center Hannover, Max Planck Institute for Informatics). The vision inspiring LivingKnowledge is to consider diversity an asset and to make it traceable, understandable, and exploitable, with the goal to improve navigation, exploration, and search in very large multimodal datasets. Our department D5 coordinates the work on knowledge evolution and

contributes to advanced fact and opinion extraction, advanced clustering and aggregation, as well as foundations and representation. Main research challenges are the effect of diversity and time on opinions and bias (e.g., in entertainment media or on political topics). Within LivingKnowledge we extend our ongoing foundational research in the area of multimodal data analysis, knowledge management, ontology evolution, and advanced Web mining.

SAPIR

SAPIR (Search on Audio-Visual Content using Peer-to-Peer Information Retrieval, `http://www.sapir.eu/`) is a STREP (Specific Targeted Research Project) that started in January 2007 and involves 9 partners from academia and industry (IBM Research Haifa, Telenor Norway, Telefonica Spain, Xerox Grenoble, Eurix Milano, CNR Pisa, University Padova, University of Brno, Max Planck Institute for Informatics). It aims to build a full-fledged search engine for multimedia content like images, videos, and multimodal information like photos that have speech or textual descriptions and user-provided tags as well as metadata, using peer-to-peer overlay networks.

Our department D5 is coordinating the work on the overall system architecture, and contributes to research on peer-to-peer indexing, query processing, and result ranking. We leverage our ongoing foundational research on query routing for peer-to-peer text search, statistics for peer-content synopses, and decentralized link analysis. We also build on our system expertise and experience with our Minerva prototype system, but we will also pursue major extensions and couple Minerva with software components provided by the project partners. The main research challenges lie in scalable query processing based on distributed indexing and caching, multimodal scoring and ranking for high-precision search results, and harnessing information about social networks.

As of March 2009, a distributed platform has been built that couples components developed by four different partners (image similarity indexes at University of Brno, text and metadata indexes at the Max Planck Institute for Informatics, query processor at IBM Haifa, front-end server at CNR Pisa), each of which is in turn built as a peer-to-peer system deployed on local computers. Our project partners have compiled a large test collection (`http://cophir.isti.cnr.it/`) with 100 million photos crawled from Flickr, including their MPEG-7 content features, metadata, as well as user-provided tags and comments. The complete SAPIR platform has been tested with this large-scale collection, and demonstrated on various occasions including the Cebit fair in March 2009 (Europe's largest trade show in the ICT area).

AEOLUS

AEOLUS (Algorithmic Principles for Building Efficient Overlay Computers, `http://aeolus.ceid.upatras.gr/`) is an Integrated Project (IP), running from September 2005 until August 2009. It involves 22, mostly academic, partners and aims at foundations for efficient cloud computing (global overlay computing) in large-scale networks. The Max Planck Institute for Informatics participates in the subproject SP3 on Sharing of Information and Computation. The emphasis of our work in this context has been on dynamic replication for scalable data management with response-time and availability guarantees. We have extended

previous models for data replication in peer-to-peer overlays and other large-scale networks, and have developed a near-optimal algorithm, coined P2R2 (peer-to-peer replication with 2-approximation), that has particularly strong capabilities for dynamic adaptation and outperforms the best previously known methods by a large margin. Our deliverables for the project include prototype software for P2R2.

DELIS

DELIS (Dynamically Evolving Large-Scale Information Systems, `http://delis.upb.de/`) was an Integrated Project (IP), running from January 2004 until February 2008. It involved 20, mostly academic, partners and aimed at foundations for the analysis and self-organization of complex systems such as peer-to-peer information sharing. DELIS was one of four EU projects funded under the Complex Systems Initiative. The Max Planck Institute for Informatics participated with two of its groups, D1 and D5. The project had, to some extent, a catalytic function in stimulating joint research between the two groups and led to a number of publications in first-rate venues.

Our department D5 coordinated one of the six subprojects of DELIS, namely, SP6 on Data Management, Search, and Mining on Internet-Scale, Dynamically Evolving Peer-to-Peer Networks. Our main contributions were major building blocks for peer-to-peer-based Web search, integrated into the comprehensive testbed Minerva with scalable methods for query routing, distributed statistics management, and decentralized link analysis. One intriguing use case for these results is the original DELIS scenario of a peer-to-peer search engine with autonomous peers, each managing and (selectively) posting its own content, and collaborative search for high-quality query results without any centralized control. Such a network of search engines would be an interesting alternative to the current oligopoly of commercial search engines, and would be more robust to conceivable distortions of search results by commercial bias or even censorship. Alternatively, many of the algorithmic techniques that we have developed in DELIS are also applicable to large server farms, the architectural basis of today's search engine technology. For example, distributed link analysis algorithms are very useful also for running authority computations and other forms of graph analyses in data centers with many commodity computers with a high-speed network.

Results of DELIS, including the Minerva prototype system, are used in the ongoing SAPIR project for distributed search of audio-visual data (see Subsection 31.13.1). There, we investigate both of the above usage scenarios, referred to as *peer-centric* and *network-centric* architectures.

Network of Excellence DELOS

DELOS was a network of excellence on digital library systems. It ran from January 2004 through December 2007, and involved about 50 partners, organized into 8 thematic clusters. D5 participated in clusters WP1 and WP2 on digital library system architecture and personalized access.

The main issues pursued by D5 in this context were query routing and query processing over federations of digital libraries and the exploitation of personal profiles for routing strategies. D5 collaborated on this task with the following universities: University of Athens, University

of Brno, University of Duisburg, ETH Zurich, and University of Basel. In addition, a related but more specific task was on analyzing access and behavior information of digital-library users. This work centered around the action logs of the TEL portal to national libraries in Europe. It involved collaboration with TEL (The European Library), the University of Padova, and the University of Athens.

31.13.2 Projects Funded by the German Science Foundation (DFG)

Participation in the Excellence Cluster on Multimodal Computing and Interaction

D5 participates in the DFG Excellence Cluster on Robust, Efficient, and Intelligent Processing of Multimodal Information (text, speech, visual data, etc.), the institute's lead theme. Gerhard Weikum is one of the cluster's 12 principal investigators. Ralf Schenkel leads an independent research group at Saarland University, which closely collaborates with us. The research area that D5 mostly focuses on is RA5 on the theme of an "Open Science Web". The vision that drives this research direction is that of a comprehensive, multimodal Open Science Web that encompasses all human knowledge in terms of concepts, entities, and semantic relations between them. The Open Science Web should enable knowledge search, exploration, and analysis in a way that significantly boosts the productivity of skilled knowledge workers compared to today's use of sources like Google, Wikipedia, or PubMed. This long-term goal entails research on large-scale information extraction, relation learning, advanced search methods, social communities, and information history mining. Our foundational work on knowledge harvesting (see Section 31.4) fits very well with this theme in the Excellence Cluster.

Network-based Interactive Navigation and Analysis of Large Biological Datasets

As part of the German Priority Program (Schwerpunktprogramm) on Scalable Visual Analytics, Our department D5 pursues a joint project with the bioinformatics and algorithmic groups of Prof. Hans-Peter Lenhof at Saarland University, Prof. Oliver Kohlbacher and Prof. Michael Kaufmann, both at the University of Tübingen. The project has started in March 2009. It aims at scalable support for data-intensive analysis and visualization of biological networks. To this end, D5 is investigating the data management layer of a comprehensive object model repository for biological networks (BN++, `http://www.bnplusplus.org/`) and versatile visualization tool suite (BiNA, `http://www.bnplusplus.org/bina/`). The technical problems addressed in this context are twofold. First, we aim to provide efficient updates for large-scale biological data repositories, with support for versioning, annotation, and provenance tracking. Second, we aim to identify suitable, generic building blocks that support the upper-level analyses and visualization of biological networks (e.g., of deregulated paths), yet can be efficiently executed directly in the data-management layer. Examples of such building blocks could be incremental shortest-path computations or the identification of bounded-size subgraphs with maximum total node weight. As the RDF data model is gaining popularity for biological data (see, e.g., `http://www.uniprot.org/`), we are considering RDF as a new basis for BN++. Consequently, we are looking into using our own engine RDF-3X, developed at department D5 of the Max Planck Institute for Informatics, as the data-management layer for this joint project.

Graduiertenkolleg "Quality Guarantees for Computer Systems"

The group has actively participated in the Saarland University's graduate studies program (Graduiertenkolleg) on "Quality Guarantees for Computer Systems". Two doctoral students received scholarships from this program.

31.13.3 Other Projects and Cooperations with Industry

With financial support by the German-Israeli Foundation for Scientific Research and Development (GIF), D5 has started collaborating with Prof. Yehoshua Sagiv of the Hebrew University at Jerusalem. The topic that we jointly addressing is New Paradigms for Knowledge Queries over Data Graphs.

In addition to the industrial partners in the EU projects, D5 maintains loose collaborations (e.g., through Master's theses, internships, joint publications) with various local and non-local companies, including Microsoft, Google, and Yahoo. Over the last two years, five graduate students of D5 have done 3-month internships at these companies. Some of these resulted in joint publications.

31.14 Publications

Books and Proceedings

[1] R. Klamma, N. Sharda, B. Fernández-Manjón, H. Kosch, and M. Spaniol, eds. *Proceedings of the 1st Workshop on Story-Telling for Educational Gaming (STEG) 2008*, Aachen, 2008, *CEUR Workshop Proceedings*, vol. 386. CEUR.

[2] M. Lux, M. Granitzer, and M. Spaniol, eds. *Multimedia Semantics - The Role of Metadata*, *Studies in Computational Intelligence*, vol. 101. Springer, Berlin, 2008.

Journal articles and book chapters

[1] S. Abiteboul, R. Agrawal, P. A. Bernstein, M. J. Carey, S. Ceri, W. B. Croft, D. J. DeWitt, M. J. Franklin, H. Garcia-Molina, D. Gawlick, J. Gray, L. M. Haas, A. Y. Halevy, J. M. Hellerstein, Y. E. Ioannidis, M. L. Kersten, M. J. Pazzani, M. Lesk, D. Maier, J. F. Naughton, H.-J. Schek, T. K. Sellis, A. Silberschatz, M. Stonebraker, R. T. Snodgrass, J. D. Ullman, G. Weikum, J. Widom, and S. B. Zdonik. The Lowell database research self-assessment. *Communications of the ACM*, 48(5):111–118, 2005.

[2] A. Ailamaki, S. Chaudhuri, S. Lightstone, G. M. Lohman, P. Martin, K. Salem, and G. Weikum. Report on the Second International Workshop on Self-Managing Database Systems (SMDB 2007). *IEEE Data Engineering Bulletin*, 30(2):2–4, 2007.

[3] S. Amer-Yahia, V. Markl, A. Y. Halevy, A. Doan, G. Alonso, D. Kossmann, and G. Weikum. Databases and Web 2.0 panel at VLDB 2007. *SIGMOD Record*, 37(1):49–52, 2008.

[4] R. Angelova, M. Lipczak, E. Milios, and P. Pralat. Investigating the properties of a social bookmarking and tagging network. *International Journal of Data Warehousing and Mining (IJDWM)*, 5, 2009.

[5] W. Bailer, L. Brunie, M. Döller, M. Granitzer, R. Klamma, H. Kosch, M. Lux, and M. Spaniol. Multimedia metadata standards. In B. Furht, ed., *Encyclopedia of Multimedia*, pp. 568–574. Springer, Berlin, 2nd edition, 2008.

[6] M. Bender, T. Crecelius, M. Kacimi, S. Michel, J. X. Parreira, and G. Weikum. Peer-to-peer information search: Semantic, social, or spiritual? *IEEE Data(base) Engineering Bulletin*, 30(2):51–60, 2007.

[7] S. Chernov, P. Serdyukov, M. Bender, S. Michel, G. Weikum, and C. Zimmer. Database selection and result merging in P2P Web search. In G. Vijay and M. R. Ponnada, eds., *Dynamics of Search Engines: An Introduction*, ch. 5, pp. 56–71. The Icfai University Press, Hyderabad, India, 2007.

[8] J. Graupmann, M. Biwer, C. Zimmer, P. Zimmer, M. Bender, M. Theobald, and G. Weikum. COMPASS: A concept-based Web search engine for HTML, XML, and Deep Web data. In G. Vijay and M. R. Ponnada, eds., *Dynamics of Search Engines: An Introduction*, ch. 12, pp. 193–203. The Icfai University Press, Hyderabad, India, 2007.

[9] L. Jäger, M. Jarke, R. Klamma, and M. Spaniol. Transkriptivität: Operative Medientheorien als Grundlage von Informationssystemen für die Kulturwissenschaften. *Informatik Spektrum*, 31(1):21–29, 2008.

[10] G. Kasneci, F. Suchanek, M. Ramanath, and G. Weikum. The YAGO-NAGA approach to knowledge discovery. *SIGMOD Record Special Issue on Managing Information Extraction*, 37(4):41–47, 2008.

[11] R. Klamma, Y. Cao, and M. Spaniol. Smart social software for mobile cross-media communities. In M. Lux, M. Granitzer, and M. Spaniol, eds., *Multimedia Semantics - The Role of Metadata*, *Studies in Computational Intelligence*, vol. 101, pp. 87–106. Springer, Berlin, 2008.

[12] T. Neumann and S. Michel. Algebraic query optimization for distributed top-k queries. *Informatik - Forschung und Entwicklung*, 21(3-4):197–211, 2007.

[13] T. Neumann and G. Weikum. RDF-3X: a RISC-style engine for RDF. *Proceedings of the VLDB Endowment*, 1(1):647–659, 2008.

[14] N. Ntarmos, P. Triantafillou, and G. Weikum. Distributed hash sketches: Scalable, efficient, and accurate cardinality estimation for distributed multisets. *ACM Transactions on Computer Systems*, 27(1):1–53, 2009.

[15] N. Ntarmos, P. Triantafillou, and G. Weikum. Statistical structures for Internet-scale data management. *The VLDB Journal*, 18, 2009. Available via Springer online-first: http://www.springerlink.com/content/a40238582233706r/.

[16] J. X. Parreira, C. Castillo, D. Donato, S. Michel, and G. Weikum. The JXP method for robust PageRank approximation in a peer-to-peer Web search network. *The VLDB Journal*, 16(3), 2007.

[17] J. X. Parreira, C. Castillo, D. Donato, S. Michel, and G. Weikum. The Juxtaposed approximate PageRank method for robust PageRank approximation in a peer-to-peer Web search network. *VLDB Journal*, 17(2):291–313, 2008.

[18] R. Schenkel, T. Crecelius, M. Kacimi, T. Neumann, J. Parreira, M. Spaniol, and G. Weikum. Social wisdom for search and recommendation. *IEEE Data Engineering Bulletin*, 31(2):40–49, 2008.

[19] M. Spaniol, R. Klamma, and Y. Cao. Media centric knowledge sharing on the Web 2.0. In M. Lytras, R. Tennyson, and P. Ordonez de Pablos, eds., *Knowledge Networks: The Social Software Perspective*, ch. 4, pp. 46–60. IGI, Hershey, 2009.

[20] M. Spaniol, R. Klamma, and M. Lux. Imagesemantics: User-generated metadata, content based retrieval & beyond. *Journal of Universal Computer Science*, 14(10):1792–1807, 2008.

[21] F. Suchanek, G. Kasneci, and G. Weikum. YAGO - a large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 6(3):203–217, 2008.

[22] M. Theobald, H. Bast, D. Majumdar, R. Schenkel, and G. Weikum. TopX: Efficient and versatile top-k query processing for semistructured data. *The VLDB Journal*, 17(2):81–115, 2008.

[23] C. Tryfonopoulos, C. Zimmer, M. Koubarakis, and G. Weikum. Architectural alternatives for information filtering in structured overlay networks. *IEEE Internet Computing*, 11(4):24–34, 2007.

[24] G. Weikum, G. Kasneci, F. Suchanek, and M. Ramanath. Database and information-retrieval methods for knowledge discovery. *Communications of the ACM*, 52(1), 2009.

Conference articles

[1] W. Allasia, F. Falchi, F. Gallo, M. Kacimi, A. Kaplan, J. Mamou, Y. Mass, and N. Orio. Audio-visual content analysis in P2P networks: The SAPIR approach. In *The 19th International Conference on Database and Expert Systems Application (DEXA 2008)*, Turin, Italy, 2008, pp. 610–614. IEEE Computer Society.

[2] A. Anand, S. Bedathur, K. Berberich, R. Schenkel, and C. Tryfonopoulos. EverLast: A distributed architecture for preserving the Web. In *Proceedings of the Joint Conference on Digital Libraries (JCDL 2009)*, Austin, Texas, 2009.

[3] R. Angelova, G. Kasneci, F. M. Suchanek, and G. Weikum. GRAFFITI: Node labeling in heterogeneous networks (poster). In *18th International World Wide Web Conference (WWW 2009)*, Madrid, Spain, 2009. ACM.

[4] R. Angelova, M. Lipczak, E. Milios, and P. Pralat. Characterizing a social bookmarking and tagging network. In *Mining Social Data Workshop, 18th European Conference on Artificial Intelligence, 2008*, Patras, Greece, 2008. IOS.

[5] I. Arikan, S. Bedathur, and K. Berberich. Time Will Tell: Leveraging temporal expressions in IR. In *Second ACM International Conference on Web Search and Data Mining (WSDM '09) - Late Breaking Results*, Barcelona, Spain, 2009. ACM.

[6] M. Backes, M. Hamerlik, A. Linari, M. Maffei, C. Tryfonopoulos, and G. Weikum. Anonymous and censorship resistant content sharing in unstructured overlays. In R. A. Bazzi and B. Patt-Shamir, eds., *Twenty-Seventh Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2008)*, Toronto, Canada, 2008, pp. 429–429. ACM.

[7] H. Bast, A. Chitea, F. Suchanek, and I. Weber. ESTER: Efficient Search on Text, Entities, and Relations. In C. Clarke, N. Fuhr, and N. Kando, eds., *30th International Conference on Research and Development in Information Retrieval (SIGIR 2007)*, Amsterdam, The Netherlands, 2007, pp. 671–678. ACM.

[8] H. Bast, F. Suchanek, and I. Weber. Semantic full-text search with ESTER: Scalable, easy, fast. In *8th IEEE International Conference on Data Mining (ICDM 2008)*, Pisa, Italy, 2008, pp. 959–962. IEEE Computer Society.

[9] M. Bender, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum. Exploiting social relations for query expansion and result ranking. In *Data Engineering for Blogs, Social Media, and Web 2.0, ICDE 2008 Workshops*, Cancun, Mexico, 2008, pp. 501–506. IEEE Computer Society.

[10] M. Bender, T. Crecelius, S. Michel, and J. X. Parreira. P2P Web search: Make it light, make it fly (demo). In *3rd Biennial Conference on Innovative Data System Research (CIDR 2007)*, Asilomar, USA, 2007, pp. 164–168. www.crdrdb.org.

[11] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. FluxCapacitor: Efficient time-travel text search. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *33rd International Conference on Very Large Databases (VLDB 2007)*, Vienna, Austria, 2007, pp. 1414–1417. ACM.

[12] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, eds., *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, Amsterdam, Netherlands, 2007, pp. 519–526. ACM.

[13] K. Berberich, S. Bedathur, M. Vazirgiannis, and G. Weikum. Comparing apples and oranges: Normalized PageRank for evolving graphs. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 1145–1146. ACM.

[14] K. Berberich, S. Bedathur, and G. Weikum. A pocket guide to Web history. In I. Ziviani and R. A. Baeza-Yates, eds., *14th String Processing and Information Retrieval Symposium (SPIRE 2007)*, Santiago, Chile, 2007, *LNCS 4726*, pp. 86–97. Springer.

[15] K. Berberich, S. Bedathur, and G. Weikum. Tunable word-level index compression for versioned corpora. In *Proceedings of the Efficiency Issues in Information Retrieval Workshop, co-located with ECIR 2008*, Glasgow, Scotland, 2008, Lecture Notes in Computer Science. Springer.

[16] A. Broschart. Efficient integration of proximity for text, semistructured and graph retrieval. In *SIGIR 2007 Doctoral Consortium*, Amsterdam, The Netherlands, 2007, p. 917. ACM.

[17] A. Broschart and R. Schenkel. Effiziente Textsuche mit Positionsinformation. In H. Höpfner and F. Klan, eds., *20. Workshop über Grundlagen von Datenbanken*, Apolda, Germany, 2008, pp. 101–105. International University in Germany.

[18] A. Broschart and R. Schenkel. Proximity-aware scoring for XML retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2008)*, New York, USA, 2008, pp. 845–846. ACM.

[19] A. Broschart, R. Schenkel, and M. Theobald. Proximity-aware scoring for XML retrieval. In S. Geva, J. Kamps, and A. Trotman, eds., *Preproceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, Schloss Dagstuhl, Germany, 2008, pp. 46–49.

[20] A. Broschart, R. Schenkel, M. Theobald, and G. Weikum. TopX @ INEX 2007. In N. Fuhr, M. Lalmas, and A. Trotman, eds., *Preproceedings of the 6th International Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2007)*, Schloss Dagstuhl, Germany, 2007, pp. 84–93. Springer.

[21] Y. Cao, A. Glukhova, R. Klamma, D. Renzel, and M. Spaniol. Measuring community satisfaction across gaming communities. In R. Lau and B. Wah, eds., *Proceedings of the International Workshop on Interactive Digital Entertainment Technologies (IDET 2008)*, Lanzhou, China, 2008, pp. 414–419. IEEE.

[22] T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum. Making sense: Socially enhanced search and exploration. In *Proceedings of the 34th International Conference on Very Large Data Bases (VLDB 2008)*, Auckland, New Zealand, 2008, vol. 1, pp. 1480–1483. ACM.

[23] T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum. Social recommenations at work (demo). In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2008)*, Singapore, Singapore, 2008, pp. 884–884. ACM.

[24] A. Das Sarma, A. Deshpande, T. Hubauer, I. F. Ilyas, B. König-Ries, M. Renz, and M. Theobald. Working group report: Lineage/provenance. In C. Koch, B. König-Ries, V. Markl, and M. van Keulen, eds., *Uncertainty Management in Information Systems*, Schloss Dagstuhl, Wadern, Germany, 2009, no. 08421 in Dagstuhl Seminar Proceedings. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[25] M. Dudev, S. Elbassuoni, J. Luxenburger, M. Ramanath, and G. Weikum. Personalizing the search for knowledge. In *2nd International Workshop on Personalized Access, Profile Management, and Context Awareness: Databases (PersDB 2008)*, Auckland, New Zealand, 2008, pp. 1–8.

[26] S. Elbassuoni, J. Luxenburger, and G. Weikum. Adaptive personalization of Web search. In K. Rodden, I. Ruthven, and R. W. White, eds., *Proceedings of the 1st Workshop on Web Information Seeking and Interaction*, Amsterdam, The Netherlands, 2007, pp. 1–5. ACM.

[27] S. Euijong Whang, D. Menestrina, G. Koutrika, M. Theobald, and H. Garcia-Molina. Entity resolution with iterative blocking. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2009)*, Providence, Rhode Island, USA, 2009. ACM.

[28] F. Falchi, M. Kacimi, Y. Mass, F. Rabitti, and P. Zezula. SAPIR: Scalable and distributed image searching. In *SAMT (Posters and Demos)*, Genoa, Italy, 2007, *CEUR Workshop Proceedings*, vol. 300, pp. 11–12. CEUR-WS.org.

[29] S. Helmer, R. Aly, T. Neumann, and G. Moerkotte. Indexing set-valued attributes with a multi-level extendible hashing scheme. In R. Wagner, N. Revell, and G. Pernul, eds., *18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, Regensburg, Germany, 2007, *LNCS 4653*, pp. 98–108. Springer.

[30] G. Ifrim and G. Weikum. Fast logistic regression for text categorization with variable-length n-grams. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2008)*, Las Vegas, Nevada, USA, 2008, pp. 354–362. ACM.

[31] M. Kacimi and K. Yetongnon. Dimensionality reduction in a P2P system. In A. M. Tjoa and R. R. Wagner, eds., *Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, Regensburg, Germany, 2007, pp. 804–808. IEEE Computer Society.

[32] N. Kapoor, G. Das, V. Hristidis, S. Sudarshan, and G. Weikum. STAR: A system for tuple and attribute ranking of query answers (demo). In *Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007)*, Istanbul, Turkey, 2007, pp. 1483–1484. IEEE Computer Society.

[33] G. Kasneci, M. Ramanath, M. Sozio, F. Suchanek, and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *Proceedings of the 25th International Conference on Data Engineering (ICDE 2009)*, Shanghai, China, 2009. IEEE Computer Society.

[34] G. Kasneci and T. Schwentick. The complexity of reasoning about pattern-based XML schemas. In L. Libkin, ed., *26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2007)*, Beijing, China, 2007, pp. 155–164. ACM.

[35] G. Kasneci, F. Suchanek, G. Ifrim, S. Elbassuoni, M. Ramanath, and G. Weikum. NAGA: Harvesting, searching and ranking knowledge (demo). In J. Tsong-Li Wang, ed., *Proceedings of the ACM SIGMOD 2008 International Conference on Management of Data (SIGMOD 2008)*, Vancouver, Canada 2008, 2008, pp. 1285–1288. ACM.

[36] G. Kasneci, F. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. In *24th International Conference on Data Engineering (ICDE 2008)*, Cancun, Mexico, 2008, pp. 953–962. IEEE Computer Society.

[37] G. Kasneci, F. Suchanek, M. Ramanath, and G. Weikum. How NAGA uncoils: Searching with entities and relations. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, New York, NY, USA, 2007, pp. 1167–1168. ACM Press.

[38] J. Luxenburger, S. Elbassuoni, and G. Weikum. Matching task profiles and user needs in personalized Web search. In *17th Conference on Information and Knowledge Management (CIKM 2008)*, Napa Valley, USA, 2008, pp. 689–698. ACM.

[39] J. Luxenburger, S. Elbassuoni, and G. Weikum. Task-aware search personalization. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, eds., *31st Annual International ACM SIGIR Conference (SIGIR 2008)*, Singapore, 2008, pp. 721–722. ACM.

[40] J. Luxenburger, E. van der Meulen, and G. Weikum. A user-interaction model for the European library portal. In *Proceedings of the 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries*, Corfu, Greece, 2007. http://www.dblab.ntua.gr/persdl2007/papers/22.pdf.

[41] J. Luxenburger and G. Pandey. Exploiting session context for information retrieval - a comparative study. In *30th European Conference on Information Retrieval (ECIR 2008)*, Glasgow, Scotland, 2008, pp. 652–657. Springer.

[42] A. Mazeika, M. H. Boehlen, and D. Trivellato. Analysis and interpretation of visual hierarchical heavy hitters of binary relations. In P. Atzeni, A. Caplinskas, and H. Jaakkola, eds., *12th East European Conference on Advances in Databases and Information Systems (ADBIS 2008)*, Pori, Finland, 2008, *LNCS 5207*, pp. 168–183. Springer.

[43] A. Mazeika, A. Taliun, and M. H. Böhlen. CORE: Nonparametric clustering of large numeric databases. In *SIAM International Conference on Data Mining (SDM 2009)*, Sparks, Nevada, 2009. Society for Industrial and Applied Mathematics.

[44] G. de Melo and S. Siersdorfer. Multilingual text classification using ontologies. In G. Amati, C. Carpineto, and G. Romano, eds., *Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007)*, Rome, Italy, 2007, *LNCS 4425*, pp. 541–548. Springer. Acceptance Ratio 1:4.

[45] G. de Melo, F. Suchanek, and A. Pease. Integrating YAGO into the Suggested Upper Merged Ontology. In *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008)*, Dayton, OH, USA, 2008, vol. 1, pp. 190–193. IEEE Computer Society.

[46] G. de Melo and G. Weikum. On the utility of automatically generated wordnets. In A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, and P. Vossen, eds., *Fourth Global WordNet Conference (GWC 2008)*, Szeged, Hungary, 2007, pp. 147–161. University of Szeged.

[47] G. de Melo and G. Weikum. Language as a foundation of the Semantic Web. In C. Bizer and A. Joshi, eds., *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008)*, Aachen, Germany, 2008, *CEUR-WS*, vol. 401, p. 401. CEUR.

[48] G. de Melo and G. Weikum. A machine learning approach to building aligned wordnets. In *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, Hong Kong, 2008, pp. 163–170.

[49] G. de Melo and G. Weikum. Mapping Roget's thesaurus and WordNet to French. In *6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco, 2008. European Language Resources Association.

[50] S. Michel and T. Neumann. Search for the best but expect the worst - distributed top-k queries over decreasing aggregated scores. In *Proceedings of the Tenth International Workshop on the Web and Databases (WebDB 2007)*, Beijing, China, 2007. ACM.

[51] G. Moerkotte and T. Neumann. Dynamic programming strikes back. In J. T.-L. Wang, ed., *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2008)*, Vancouver, Canada, 2008, pp. 539–552. ACM.

[52] G. Moerkotte and T. Neumann. Faster join enumeration for complex queries. In *Proceedings of the 24th International Conference on Data Engineering (ICDE 2008)*, Washington, USA, 2008, pp. 1430–1432. IEEE.

[53] P. Moreno-Ger, M. Spaniol, E. López Manas, N. Drobek, and B. Fernández-Manjón. Making sense of collaboratively annotated multimedia metadata for (mobile) digital story-telling and educational gaming. In R. Klamma, N. Sharda, B. Fernández-Manjón, H. Kosch, and M. Spaniol, eds., *Proceedings of the 1st Workshop on Story-Telling for Educational Gaming (STEG) 2008*, Maastricht, The Netherlands, 2008, *CEUR Workshop Proceedings*, vol. 386. CEUR.

[54] T. Neumann. Optimizing ranked retrieval. In W. Wagner, N. Revell, and G. Pernul, eds., *18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, Regensburg, Germany, 2007, *LNCS 4653*, pp. 329–338. Springer.

[55] T. Neumann. Query simplification: Graceful degradation for join-order optimization. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD 2009)*, Providence, USA, 2009. ACM.

[56] T. Neumann, M. Bender, S. Michel, R. Schenkel, P. Triantafillou, and G. Weikum. Optimizing distributed top-k queries. In *Proceedings of the 9th International Conference on Web Information Systems (WISE 2008)*, Auckland, New Zealand, 2008, *LNCS 5175*, pp. 337–349. Springer.

[57] T. Neumann and S. Michel. Algebraic query optimization for distributed top-k queries. In A. Kemper, H. Schöning, T. Rose, M. Jarke, T. Seidl, C. Quix, and C. Brochhaus, eds., *12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web*, Aachen, Germany, 2007, *Lecture Notes in Informatics*, vol. 103, pp. 324–343. Gesellschaft für Informatik.

[58] T. Neumann and S. Michel. Smooth interpolating histograms with error guarantees. In W. A. Gray, K. G. Jeffery, and J. Shao, eds., *Sharing Data, Information and Knowledge, 25th British National Conference on Databases, BNCOD 25*, Cardiff, UK, 2008, *LNCS 5071*, pp. 126–138. Springer.

[59] T. Neumann and G. Moerkotte. A framework for reasoning about share equivalence and its integration into a plan generator. In *Datenbanksysteme in Business, Technologie und Web (BTW 2009), 13. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS)*, Münster, Germany, 2009, Lecture Notes in Informatics. GI.

[60] T. Neumann and G. Weikum. Scalable join processing on very large RDF graphs. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD 2009)*, Providence, USA, 2009. ACM.

[61] H. Pan, R. Schenkel, and G. Weikum. Fine-grained relevance feedback for XML retrieval (demo). In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2008)*, Singapore, Singapore, 2008, pp. 887–887. ACM.

[62] J. X. Parreira, D. Donato, C. Castillo, and G. Weikum. Computing trusted authority scores in peer-to-peer Web search networks. In C. Castillo, K. Chellapilla, and B. Davison, eds., *Adversarial Information Retrieval on the Web (AIRWeb 2007)*, Banff, Canada, 2007, pp. 73–80. ACM.

[63] J. X. Parreira, S. Michel, M. Bender, T. Crecelius, and G. Weikum. P2P authority analysis for social communities. In C. Koch, J. Gehrke, M. N. Garofalakis, D. Srivastava, K. Aberer, A. Deshpande, D. Florescu, C. Y. Chan, V. Ganti, C.-C. Kanne, W. Klas, and E. J. Neuhold, eds., *33rd International Conference on Very Large Data Bases (VLDB 2007)*, Vienna, Austria, 2007, pp. 1398–1401. ACM.

[64] P. Raftopoulou, E. Petrakis, C. Tryfonopoulos, and G. Weikum. Information retrieval and filtering over self-organising digital libraries. In D. Christensen-Dalsgaard, Birte Castelli, B. A. Jurik, and J. Lippincott, eds., *12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*, Aarhus, Denmark, 2008, pp. 320–333. Springer.

[65] M. Ramanath and K. Sarath Kumar. A rank-rewrite framework for summarizing XML documents. In *2nd International Workshop on Ranking in Databases (DBRank), ICDE 2008 Workshops*, Cancun, Mexico, 2008, pp. 540–547. IEEE Computer Society.

[66] M. Ramanath and K. Sarath Kumar. Xoom: A tool for zooming in and out of XML documents. In *Proceedings of the 12th International Conference on Extending Database Technology (EDBT 2009)*, St. Petersburg, Russia, 2009.

[67] D. Renzel, R. Klamma, and M. Spaniol. MobSOS - a testbed for mobile multimedia community services. In *Proceedings of the Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)*, Klagenfurt, Austria, 2008, pp. 139–142. IEEE.

[68] R. Schenkel, A. Broschart, S. Hwang, M. Theobald, and G. Weikum. Efficient text proximity search. In N. Ziviani and R. A. Baeza-Yates, eds., *14th String Processing and Information Retrieval Symposium (SPIRE 2007)*, Santiago, Chile, 2007, *LNCS 4726*, pp. 287–299. Springer.

[69] R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, and G. Weikum. Efficient top-k querying over social-tagging networks. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2008)*, Singapore, Singapore, 2008, pp. 523–530. ACM.

[70] R. Schenkel, F. Suchanek, and G. Kasneci. YAWN: A semantically annotated Wikipedia XML corpus. In A. Kemper, H. Schöning, T. Rose, M. Jarke, T. Seidl, C. Quix, and C. Brochhaus, eds., *12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2007)*, Aachen, Germany, 2007, *Lecture Notes in Informatics*, vol. 103, pp. 277–291. Gesellschaft für Informatik.

[71] G. Shegalov and G. Weikum. Formal verification of a transactional interaction contract. In *IEEE SOA Industry Summit (SOAIS 2008)*, Honolulu, Hawaii, USA, 2008, pp. 525–528. IEEE Computer Society.

[72] G. Shegalov and G. Weikum. Formal verification of Web service interaction contracts. In *2008 IEEE International Conference on Services Computing (SCC 2008)*, Honolulu, Hawai, USA, 2008, pp. 525–528. IEEE Computer Society.

[73] M. Shmueli-Scheuer, C. Li, Y. Mass, H. Roitman, R. Schenkel, and G. Weikum. Best effort top-k query processing under budgetary constraints. In Y. Ioannidis, D. Lee, and R. Ng, eds., *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE 2009)*, Shanghai, China, 2009.

[74] M. Sozio, T. Crecelius, J. Xavier Parreira, and G. Weikum. Good guys vs. bad guys: Countering cheating in peer-to-peer authority computations over social networks. In *11th International Workshop on the Web and Databases (WebDB 2008)*, Vancouver, Canada, 2008, pp. 103–108. ACM.

[75] M. Sozio, T. Neumann, and G. Weikum. Near-optimal dynamic replication in unstructured peer-to-peer networks. In M. Lenzerini and D. Lembo, eds., *Proceedings of the Conference on Principles of Database System (PODS 2008)*, Vancouver, Canada, 2008, pp. 281–290. ACM.

[76] M. Spaniol, Y. Cao, R. Klamma, P. Moreno-Ger, B. Fernández-Manjón, J. L. Sierra, and G. Toubekis. From story-telling to educational gaming: The Bamiyan Valley case. In F. Li, J. Zhao, T. Shih, R. Lau, Q. Li, and D. McLeod, eds., *Advances in Web Based Learning - ICWL 2008*, Jinhua, China, 2008, *LNCS 5145*, pp. 253–264. Springer.

[77] M. Spaniol, R. Klamma, and Y. Cao. Learning as a service: A Web-based learning framework for communities of professionals on the Web 2.0. In H. Leung, F. Li, R. Lau, and Q. Li, eds., *Advances in Web Based Learning - ICWL 2007*, Edinburgh, Scotland, 2008, *LNCS 4823*, pp. 160–173. Springer.

[78] J. Stoyanovich, S. Bedathur, K. Berberich, and G. Weikum. EntityAuthority: Semantically enriched graph-based authority propagation. In *Proceedings of the 10th International Workshop on Web and Databases (WebDB 2007)*, Beijing, China, 2007.

[79] F. Suchanek. Social tags: Meaning and suggestions. In *ACM Conference on Information and Knowledge Management (CIKM 2008)*, Napa, CA, USA, 2008, pp. 223–232. ACM.

[80] F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 697–706. ACM.

[81] F. Suchanek, M. Sozio, and G. Weikum. SOFIE: A self-organizing framework for information extraction. In *Proceedings of the 18th World Wide Web Conference (WWW 2009)*, Madrid, Spain, 2009. ACM.

[82] B. Taneva, J. Giesen, P. Zolliker, and K. Mueller. Choice based conjoint analysis: Discrete choice models vs. direct regression. In E. Hüllermeier and J. Fürnkranz, eds., *Proceedings of the ECML/PKDD Workshop on Preference Learning (PL-08)*, Antwerp, Belgium, 2008, pp. 67–81.

[83] M. Theobald, M. AbuJarour, and R. Schenkel. TopX 2.0 at the INEX 2008 Efficiency Track. In S. Geva, J. Kamps, and A. Trotman, eds., *Preproceedings of the 7th Int. Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, Schloss Dagstuhl, Germany, 2008, pp. 230–244.

[84] M. Theobald, A. Broschart, R. Schenkel, S. Solomon, and G. Weikum. TopX - adhoc track and feedback task. In N. Fuhr, M. Lalmas, and A. Trotman, eds., *Comparative Evaluation of XML Information Retrieval Systems, 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, Schloss Dagstuhl, Germany, 2007, *LNCS 4518*, pp. 233–242. Springer.

[85] M. Theobald and R. Schenkel. Overview of the INEX 2008 Efficiency Track. In S. Geva, K. Kamps, and A. Trotman, eds., *Preproceedings of the 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, Schloss Dagstuhl, Germany, 2008, pp. 208–219.

[86] M. Theobald, R. Schenkel, and G. Weikum. TopX - efficient and versatile top-k query processing for text, semistructured, and structured data. In A. Kemper, H. Schöning, T. Rose, M. Jarke, T. Seidl, C. Quix, and C. Brochhaus, eds., *12. GI-Fachtagung Datenbanksysteme in Business, Technologie und Web (BTW 2007)*, Aachen, Germany, 2007, *Lecture Notes in Informatics*, vol. 103, pp. 475–485. Gesellschaft für Informatik.

[87] M. Theobald, R. Schenkel, and G. Weikum. The TopX DB&IR engine (demo). In N. Koudas, ed., *2007 ACM SIGMOD International Conference on Management of Data*, Beijing, 2007, pp. 1141–1143. ACM.

[88] M. Theobald, N. Shah, and J. Shrager. Extraction of conditional probabilities of the relationships between drugs, diseases, and genes from PubMed guided by relationships in PharmGKB. In *2009 AMIA Summit on Translational Bioinformatics*, Grand Hyatt, San Francisco, 2009. AMIA.org.

[89] A. Vlachou, M. Vazirgiannis, and K. Berberich. Representing and quantifying rank - change for the Web graph. In *Algorithms and Models for the Web-Graph, Fourth International Workshop, WAW 2006*, Banff, Canada, 2006, *LNCS 4936*, pp. 157–165. Springer.

[90] G. Weikum. DB&IR: both sides now. In C. Y. Chan, B. C. Ooi, and A. Zhou, eds., *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2007)*, Beijing, China, 2007, pp. 25–30. ACM.

[91] G. Weikum. Harvesting and organizing knowledge from the Web. In Y. E. Ioannidis, B. Novikov, and B. Rachev, eds., *Advances in Databases and Information Systems: 11th East European Conference (ADBIS 2007)*, Varna, Bulgaria, 2007, *LNCS 4690*, pp. 12–13. Springer.

[92] J. Xavier Parreira, S. Michel, and G. Weikum. Efficiently handling dynamics in distributed link based authority analysis. In J. Bailey, D. Maier, K.-D. Schewe, B. Thalheim, and X. S. Wang, eds., *9th International Conference on Web Information Systems Engineering (WISE)*, Auckland, New Zealand, 2008, pp. 36–49. Springer.

[93] Q. Zhang, F. Suchanek, and G. Weikum. TOB: Timely ontologies for business relations. In *11th International Workshop on Web and Databases 2008 (WebDB 2008)*, Vancouver, Canada, 2008. ACM.

[94] C. Zimmer, S. Bedathur, and G. Weikum. Standing on the shoulders of peers: Caching in peer-to-peer information retrieval. In *Fifth International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P)*, Vienna, Austria, 2007.

[95] C. Zimmer, S. Bedathur, and G. Weikum. Flood little, cache more: Effective result-reuse in P2P IR systems. In J. R. Haritsa, K. Ramamohanarao, and V. Pudi, eds., *13th International Conference on Database Systems for Advanced Applications (DASFAA)*, New Delhi, India, 2008, *LNCS 4947*, pp. 235–250. Spinger.

[96] C. Zimmer, J. Heinz, C. Tryfonopoulos, and G. Weikum. P2P information retrieval and filtering with MAPS (demo). In K. Wehrle, W. Kellerer, S. K. Singhal, and R. Steinmetz, eds., *Proceedings of the Eighth International Conference on Peer-to-Peer Computing (P2P 2008)*, Aachen, Germany, 2008, pp. 84–85. IEEE Computer Society.

[97] C. Zimmer, C. Tryfonopoulos, K. Berberich, M. Koubarakis, and G. Weikum. Node behavior prediction for large-scale approximate information filtering. In *1st International Workshop on Large Scale Distributed Systems for Information Retrieval (LSDS-IR 2007)*, Amsterdam, The Netherlands, 2007.

[98] C. Zimmer, C. Tryfonopoulos, K. Berberich, M. Koubarakis, and G. Weikum. Approximate information filtering in peer-to-peer networks. In J. Bailey, D. Maier, K.-D. Schewe, B. Thalheim, and X. Sean Wang, eds., *Proceedings of the 9th International Conference on Web Information Systems Engineering (WISE 2008)*, Auckland, New Zealand, 2008, *LNCS 5175*, pp. 6–19. Springer.

[99] C. Zimmer, C. Tryfonopoulos, and G. Weikum. Efficient search and approximate information filtering in a distributed peer-to-peer environment of digital libraries. In C. Thanos, F. Borri, and L. Candela, eds., *Digital Libraries: Research and Development, First International DELOS Conference*, Pisa, Italy, 2007, *LNCS 4877*, pp. 328–337. Springer.

[100] C. Zimmer, C. Tryfonopoulos, and G. Weikum. MinervaDL: An architecture for information retrieval and filtering in distributed digital libraries. In L. Kovács, N. Fuhr, and C. Meghini, eds., *Research and Advanced Technology for Digital Libraries, 11th European Conference, ECDL 2007*, Budapest, Hungary, 2007, *LNCS 4675*, pp. 148–160. Springer.

[101] C. Zimmer, C. Tryfonopoulos, and G. Weikum. Exploiting correlated keywords to improve approximate information filtering. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, eds., *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, 2008, pp. 323–330. ACM.

Theses

[1] G. Aalam. Georeferenzierung von Suchergebnissen auf Basis von annotierten Daten und Geoinformationssystemen. Bachelor thesis, Universität des Saarlandes, 2007.

[2] M. AbuJarour. Efficient XML query processing and full-text search. Masters thesis, Universität des Saarlandes, 2008.

[3] A. Anand. Index partitioning strategies for peer-to-peer Web archival. Masters thesis, Universität des Saarlandes, 2009.

[4] L. M. Andreescu. Pricing information goods in an agent-based information filtering system. Masters thesis, Universität des Saarlandes, 2008.

[5] I. Arikan. Exploiting temporal references in text retrieval. Masters thesis, Universität des Saarlandes, 2009.

[6] M. Bender. *Advanced Methods for Query Routing in Peer-to-Peer Information Retrieval*. Phd thesis, Universität des Saarlandes, 2007.

[7] T. Benke. Semantic overlay networks for peer-to-peer Web search. Bachelor thesis, Universität des Saarlandes, 2007.

[8] D. Dahrendorf. Analysis of Different Replication and Caching Strategies in the P2P Search Engine Minerva. Bachelor thesis, Universität des Saarlandes, 2008.

[9] M. Dudev. Personalization of search on structured data. Masters thesis, Universität des Saarlandes, 2008.

[10] S. Elbassuoni. Adaptive personalization of Web search. Masters thesis, Universität des Saarlandes, 2007.

[11] C. Gherbaoui. Similarity measures for query expansion in TopX. Diploma thesis, Universität des Saarlandes, 2008.

[12] J. Heinz. Implementation of an Approximate Information Filtering Approach (MAPS). Bachelor thesis, Universität des Saarlandes, 2008.

[13] S. Holder. Replication in unstructured peer-to-peer networks with availability constraints. Masters thesis, Universität des Saarlandes, 2008.

[14] M. Huester. Formale Beschreibung von Fertigungsprozessen mit der Prozessalgebra CCS. Bachelor thesis, Universität des Saarlandes, 2007.

[15] G. Ifrim. *A Statistical Learning Approach to Concept-Based Document Classification.* Phd thesis, Universität des Saarlandes, 2009.

[16] L. Kasradze. Implementation of a file-based indexing framework for the TopX search engine. Masters thesis, Universität des Saarlandes, 2008.

[17] A. Kopp. Design and implementation of an automatic semantic annotation service. Diploma thesis, Universität des Saarlandes, 2007.

[18] C. Langner. Text-based approaches for content-based image retrieval in a P2P network. Diploma thesis, Universität des Saarlandes, 2008.

[19] J. Luxenburger. *Modeling and Exploiting User Search Behavior for Information Retrieval.* Phd thesis, Universität des Saarlandes, 2008.

[20] F. Makari Manshadi. Fast distributed algorithm for replication in unstructured P2P networks. Bachelor thesis, Universität des Saarlandes, 2008.

[21] G. Manolache. Index-based snippet generation. Masters thesis, Universität des Saarlandes, 2008.

[22] S. Michel. *Top-k Aggregation Queries in Large-Scale Distributed Systems.* Phd thesis, Universität des Saarlandes, 2007.

[23] J. Müller. Komplexe Filterprädikate in RDF-3X. Bachelor thesis, Universität des Saarlandes, 2009.

[24] M. Nicolay. Design und Implementierung eines Sync-ML-Clients für das c'man-Framework für Symbian-Geräte. Bachelor thesis, Universität des Saarlandes, 2007.

[25] G. Pandey. Relevance feedback using query logs. Masters thesis, Universität des Saarlandes, 2007.

[26] A. Prohaska. Building Web query subsumption hierarchies. Masters thesis, Universität des Saarlandes, 2008.

[27] C. Raileanu. Entity extraction and contextual rule selection in the medical domain. Masters thesis, Universität des Saarlandes, 2008.

[28] O. Sammodi. Incremental relevance feedback for TopX. Masters thesis, Universität des Saarlandes, 2008.

[29] R. Socher. A learning-based hierarchical model for vessel segmentation. Masters thesis, Universität des Saarlandes, 2008.

[30] S. Solomon. Evaluation of relevance feedback algorithms for XML retrieval. Masters thesis, Universität des Saarlandes, 2007.

[31] M. Stadtmüller. An XPath 2.0 full-text query parser for the TopX search engine. Bachelor thesis, Universität des Saarlandes, 2007.

[32] F. Suchanek. *Automated Construction and Growth of a Large Ontology.* Phd thesis, Universität des Saarlandes, 2008.

[33] C. Zimmer. *Approximate Information Filtering in Structured Peer-to-Peer Networks.* Phd thesis, Universität des Saarlandes, 2008.

Technical reports

[1] S. Bedathur, K. Berberich, J. Dittrich, N. Mamoulis, and G. Weikum. Scalable phrase mining for ad-hoc text analytics. Research Report MPI-I-2009-5-006, Max-Planck-Institut for Informatics, Saarbrücken, Germany, 2009.

[2] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. Research Report MPII-I-2007-5-02, Max-Planck-Institut for Informatics, Saarbrücken, Germany, 2007.

[3] G. Kasneci, M. Ramanath, M. Sozio, F. Suchanek, and G. Weikum. STAR: Steiner tree aproximation in relationship-graphs. Research Report MPI-I-2008-5-001, MPII, Saarbrücken, Germany, 2008.

[4] G. Kasneci, F. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. Research Report MPI-I-2007-5-001, Max Planck Institute for Informatics, Saarbrücken, Germany, 2007.

[5] G. de Melo, F. Suchanek, and A. Pease. Integrating YAGO into the Suggested Upper Merged Ontology. In *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008)*, Dayton, OH, USA, 2008, vol. 1, pp. 190–193. IEEE Computer Society.

[6] T. Neumann and G. Moerkotte. Single phase construction of optimal dag-structured qeps. Research Report MPI-I-2008-5-002, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, 2008.

[7] T. Neumann and G. Weikum. The RDF-3X engine for scalable management of RDF data. Research Report MPI-I-2009-5-003, Max-Planck-Institut für Informatik, Saarbrücken, Germany, 2009.

[8] N. Preda, F. Suchanek, G. Kasneci, T. Neumann, and G. Weikum. Coupling knowledge bases and Web services for active knowledge. Research Report MPI-I-2009-5-004, Max Planck Institute for Informatics, Saarbrücken, Germany, 2009.

[9] F. Suchanek, M. Sozio, and G. Weikum. SOFIE: A self-organizing framework for information extraction. Technical Report MPI-I-2008-5-004, Max-Planck-Institut für Informatik, Saarbrücken, Germany, 2008.

[10] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago : a large ontology from wikipedia and wordnet. Research Report MPI-I-2007-5-003, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, 2007.

32 The Automation of Logic Group (RG1)

32.1 Personnel

Head of Group
Prof. Dr. Christoph Weidenbach

Researchers
Dr. Jörn Freiheit (–October 2008)
Dr. Thomas Hillenbrand
Priv.-Doz. Dr. Viorica Sofronie-Stokkermans
Dr. Duc-Khanh Tran (October 2007–September 2008)
Dr. Uwe Waldmann

PhD Students
Arnaud Fietzke (November 2007–)
Willem Hagemann (April 2008–)
Matthias Horbach
Carsten Ihlemann
Swen Jacobs
Evgeny Kruglov (March 2008–)
Manuel Lamotte-Schubert
Tianxiang Lu (March 2009–)
Patrick Wischnewski (November 2007–)

Secretaries
Roxane Wetzel (–November 2008)
Jennifer Müller (November 2008–)

32.2 Visitors

In the time period from April 2007 to April 2009, the following researchers visited our group:

Florent Jacquemard	19.04.07	CNRS and ENS Cachan
Tal Lev-Ami	26.04.07	Tel Aviv University
	08.07.07–15.07.07	
Duc-Khanh Tran	05.06.07	INRIA Nancy
Peter Baumgartner	11.07.07	Universität Koblenz
Rick Rashid	12.07.07	Microsoft Research

Andre Hagehuelsmann	12.07.07	Microsoft International
Alexander Brändle	12.07.07	Microsoft Research
Michael Rusinovitch	15.07.07	LORIA Nancy
Thomas Sturm	18.01.08	Universität Passau
	03.03.08–13.03.08	
Fabrice Nahon	08.02.08	INRIA/LORIA
Manfred Jäger	22.04.08	Aalborg University
Colas Le Guernic	26.06.08–27.06.08	Verimac, Gières
Christopher Lynch	19.07.08–08.08.08	Clarkson University
Geoff Sutcliffe	30.09.08–	University of Miami

32.3 First-order Theorem Proving

Our research in first-order theorem proving further develops the superposition framework towards a general methodology that can both be used to understand first-order reasoning in general and be appropriately instantiated to particular reasoning modes for specific first-order (sub)classes and combinations.

Case analysis, called splitting, is a powerful reasoning technique. It becomes quite involved when combined with superposition in general. In particular, the notions of fairness and redundancy need careful attention (Section 32.3.1). Calculi like model evolution with a weaker redundancy notion than superposition have more flexible and powerful splitting concepts, so there is the natural question of a trade off (Section 32.3.2) between the splitting concept and redundancy notion. Completion has a proof-theoretic redundancy notion and we have shown how this can be transferred to the superposition context (Section 32.3.3).

In order to fit more specific first-order reasoning contexts we started looking at first-order reasoning for large finite domain theories (Section 32.3.4), an ordering extension of the KBO suited for reasoning in certain theory contexts (Section 32.3.5), and a first robust and practically useful variant of contextual rewriting (Section 32.3.6).

32.3.1 Labelled Splitting

Investigators: Arnaud Fietzke and Christoph Weidenbach

The splitting rule in superposition calculi performs a case analysis on variable-disjoint components of a clause, and has been implemented in SPASS for a long time [4]. Splitting preserves satisfiability in the sense that a clause set N is satisfiable if and only if at least one of the two clause sets N_1, N_2 obtained from N by adding the respective clause components, is satisfiable. The motivation for introducing the splitting rule is that, by choosing carefully which clause to split, the resulting sets N_1 and N_2 can be easier to handle. For example, splitting into non-trivial components reduces the number of different variables in a clause, which often makes it an indispensable ingredient of superposition-based decision procedures. Furthermore, by requiring both components to contain at least one positive literal, N_1 and N_2 are closer to Horn than N. For Horn clause sets, decidability results are typically easier to establish and more efficient algorithms exist.

The formal presentation of splitting has been traditionally in terms of trees of clause sets, spanned by applications of the splitting rule. In this static view, the clause set at the root of a given tree is satisfiable if and only if at least one of the leaves is satisfiable. In a prover like SPASS, however, proof search with splitting corresponds to a depth-first search through a tree whose structure changes dynamically as parts of it are pruned during backtracking. These dynamic aspects are not represented in the static view, although they are crucial for correctness. In [1, 2], we formalized them by defining a sound and complete superposition calculus with explicit splitting and backtracking, on the basis of labelled clauses. For the completeness proof, we introduced a new notion of fairness for infinite derivations. The backtracking rule in the calculus improves on the one previously implemented in SPASS, by performing a more advanced dependency analysis on closed branches, thus pruning larger parts of the search space. This improvement is reflected in our experimental results on the TPTP problem library [3] where SPASS with improved backtracking solved 20 more problems than without. The new backtracking rule is implemented in SPASS v3.2.

References

- [1] A. L. Fietzke and C. Weidenbach. Labelled splitting. In *IJCAR*, Sydney, Australia, 2008, *LNCS 5195*, pp. 459–474. Springer.
- [2] A. L. Fietzke and C. Weidenbach. Labelled splitting. Research Report MPI-I-2008-RG1-001, MPI-INF RG1, MPI für Informatik, Campus E 1 4, 66123 Saarbrücken, 2008.

[3] G. Sutcliffe and C. B. Suttner. The TPTP problem library – CNF release v1.2.1. *Journal of Automated Reasoning*, 21(2):177–203, 1998.

- [4] C. Weidenbach, R. Schmidt, T. Hillenbrand, R. Rusev, and D. Topic. System description: Spass version 3.0. In F. Pfenning, ed., *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007, *LNAI 4603*, pp. 514–520. Springer.

32.3.2 Superposition and Model Evolution Combined

Investigators: Uwe Waldmann in cooperation with Peter Baumgartner (NICTA, Canberra, Australia)

We have developed a new calculus for first-order theorem proving with equality, which generalizes both the Superposition calculus and the Model Evolution calculus (with equality) [1]. It integrates the inference rules of Superposition and of Model Evolution while preserving the individual semantically-based redundancy criteria. The inference rules are controlled by a rather flexible selection function on positive literals. This allows non-trivial combinations where inference rule applicability is disjoint, but pure forms of both calculi can be (trivially) configured, too.

On a research-methodological level, this is an initial investigation into bridging the gap between instance-based methods and Resolution methods. Both methods are rather successful, for instance in terms of performance of implemented systems at the annual CASC theorem proving competition. However, they currently stand rather separated. They provide decision procedure for different sub-classes of first-order logic, and their inference rules are incompatible, too. For instance, subsumption deletion cannot be used to such an extent with instance-based methods as with resolution methods. Our calculus can be seen as an extension of Model

Evolution by Superposition inference rules, which can be interesting, e.g., to extend results for Superposition as a decision procedure beyond currently known fragments. Alternatively, it can be seen to extend Superposition with a new splitting rule that allows, as a special case, to split a clause into *non* variable disjoint subclauses, which is interesting, e.g., to obtain a decision procedure for function-free clause logic.

References

- [1] P. Baumgartner and U. Waldmann. Superposition and model evolution combined. In R. Schmidt, ed., *22nd International Conference on Automated Deduction, CADE-22*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer.

32.3.3 Redundancy in Completion and Superposition

Investigator: Thomas Hillenbrand

As is known, the superposition calculus combines techniques of ordered resolution on the clause level with such of Knuth-Bendix completion on the term level. The WALDMEISTER system [4] is an implementation of the latter and features a simple, but effective deletion rule for theories with operators that are associative and commutative. Since its publication in [2] and [1], this deletion rule has spread to the first-order provers E [6] and PROVER9 [5], which are based on variants of superposition. Interestingly, a correctness proof has been missing so far [7].

Upon inspection [3, Chap. 7], we could demonstrate that in the superposition framework, the deletion rule is in fact not correct with respect to the standard notion of redundancy. Even worse, in every previous correctness proofs for the completion setting we found a gap. Searching for a more abstract justification, we have been able to show that ground confluence of rewrite systems can be rephrased as a property of proofs, namely that rewrite proofs of ground instances are essentially ground instances of first-order level proofs. Thanks to this observation, the deletion rule now generalizes to any algebraic structure with ground convergent rewrite system, like Abelian groups or rings. Returning to superposition, correctness does carry over with a refined literal ordering, which even can be extended such that superposition redundancy subsumes completion redundancy.

References

[1] J. Avenhaus, T. Hillenbrand, and B. Löchner. On using ground joinable equations in equational theorem proving. *Journal of Symbolic Computation*, 36(1-2):217–233, 2003.

[2] T. Hillenbrand. Schnelles Gleichheitsbeweisen: Vom Vervollständigungskalkül zum WALDMEISTER-System. Diplomarbeit, Universität Kaiserslautern, Fachbereich Informatik, 2000.

- [3] T. Hillenbrand. *Superposition and Decision Procedures – Back and Forth.* Phd thesis, Universität des Saarlandes, 2008.

[4] B. Löchner and T. Hillenbrand. A phytography of WALDMEISTER. *AI Communications*, 15(2-3):127–133, 2002.

[5] W. McCune. PROVER9 manual, 2008. Available via `http://www.prover9.org`.

[6] S. Schulz. System abstract: E 0.61. In R. Goré, A. Leitsch, and T. Nipkow, eds., *Proceedings of the First International Joint Conference on Automated Reasoning*, 2001, *LNAI 2083*, pp. 370–375. Springer-Verlag.

[7] S. Schulz. Personal communication, Dagstuhl, 2007.

32.3.4 Reasoning in Large Ontologies

Investigators: Christoph Weidenbach and Patrick Wischnewski

The goal of this work is to use a first-order theorem prover, in particular SPASS [8], to reason about huge sets of clauses over a finite domain. A typical application is to reason about knowledge comprised in an ontology. We will consider the YAGO ontology [7] which consists of 15 million facts and 1.7 million entities.

The YAGO ontology may become inconsistent while new facts are automatically extracted from the Internet (Wikipedia). Therefore, during extraction a confidence measure is added to each fact in order to decide on a consistent subset with "maximal" confidence later on. We are currently analyzing various logics that are able to identify a minimal contradiction with respect to a degree of confidence by methods compatible with resolution-style reasoning. Therefore, *Bayesian network* and *Markov network* approaches [6] do not apply whereas *Fuzzy Logic* [5], *Possibilistic Logic* [2] and *Many-Valued Logic* [4, 1, 3] are compatible with resolution-style reasoning.

Querying the YAGO ontology is based on methods used in information retrieval. We are going to define a logically well-founded framework for querying and reasoning about the knowledge contained in YAGO.

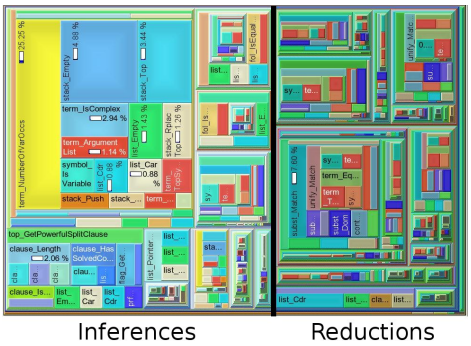


Figure 32.1: Resource consumption of SPASS

Figure 32.1 shows the resource consumption of a typical SPASS run saturating a small 1 MB portion of YAGO. The left side of Figure 32.1 depicts the resources needed for inferences and the right side the resources needed for the reduction. The run differs significantly from runs on problems coming, e.g., from verification applications. For verification applications resources for computing inferences can almost be neglected. This means that we have to investigate new technology and a different prover architecture in order to cope with YAGO.

References

[1] M. Baaz and C. G. Fermüller. Resolution-based theorem proving for manyvalued logics. *J. Symbolic Computation*, 19(4):353–391, 1995.

[2] D. Dubois and H. Prade. Resolution principles in possibilistic logic. *Int. J. Approx. Reasoning*, 4(1):1–21, 1990.

[3] H. Ganzinger and V. Sofronie-Stokkermans. Chaining techniques for automated theorem proving in many-valued logics. In *Proceedings of the 30th IEEE International Symposium on Multiple-Valued Logic (ISMVL-00)*, Portland, Oregon, 2000, pp. 337–344. IEEE.

[4] R. Hähnle and G. Escalada-Imaz. Deduction in many-valued logics: a survey. In *Mathware & Soft Computing, iv(2):69–97*, 1997.

[5] P. Hájek. *Metamathematics of Fuzzy Logic (Trends in Logic)*. Springer, 1998.

[6] J. Pearl. *Probabilistic reasoning in intelligent systems*. The Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, rev. 2nd edition, 1988.

- [7] F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In C. L. Williamson, M. E. Zurko, and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 697–706. ACM.

- [8] C. Weidenbach, R. Schmidt, T. Hillenbrand, R. Rusev, and D. Topic. System description: Spass version 3.0. In F. Pfenning, ed., *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007, *LNAI 4603*, pp. 514–520. Springer.

32.3.5 Transfinite Knuth-Bendix Ordering

Investigators: Michel Ludwig and Uwe Waldmann

In theorem proving calculi like superposition, reduction orderings are crucial to reduce the search space. Among these orderings, the Knuth-Bendix ordering is usually preferred in state-of-the-art implementations of theorem provers, as it can be efficiently implemented and as it correlates well with the sizes of terms; so, reductions w.r.t. a KBO usually lead to terms with fewer nodes. On the other hand, it is exactly this correlation between the KBO and term sizes that renders the KBO incompatible with special requirements occurring in certain applications. One example is hierarchic theorem proving [2], where one considers two signatures $\Sigma \supseteq \Sigma_0$ and needs an ordering in which every ground term involving a symbol from $\Sigma \setminus \Sigma_0$ is larger than every ground term over Σ_0. A second example is definitions of the form $f(t_1, \ldots, t_n) \approx t_0$ where f does not occur in t_0. If some variable occurs more often in t_0 than in $f(t_1, \ldots, t_n)$, such a definition cannot be handled adequately using a KBO.

In (Ludwig and Waldmann [3] we present a variant of the Knuth-Bendix ordering that uses certain ordinal numbers as weights and additionally includes a multiplicative component to cater for non-linearity. The ordering preserves as much as possible of the spirit of KBO (e.g., it is a simplification ordering that can be made total on ground terms), yet satisfies the requirements for hierarchic theorem proving and non-linear definitions. A special case of this ordering has been integrated into the MetiTarski prover by Akbarpour and Paulson [1], leading typically to a doubling in speed (Paulson, personal communication) on their benchmark tests.

References

[1] B. Akbarpour and L. C. Paulson. MetiTarski: An automatic prover for the elementary functions. In S. Autexier, J. Campbell, J. Rubio, V. Sorge, M. Suzuki, and F. Wiedijk, eds., *Intelligent Computer Mathematics, 9th International Conference, AISC 2008, 15th Symposium, Calculemus 2008, 7th International Conference, MKM 2008*, 2008, *LNCS 5144*, pp. 217–231. Springer.

[2] H. Ganzinger, V. Sofronie-Stokkermans, and U. Waldmann. Modular proof systems for partial functions with Evans equality. *Information and Computation*, 204(10):1453–1492, 2006.

• [3] M. Ludwig and U. Waldmann. An extension of the Knuth-Bendix ordering with LPO-like properties. In N. Dershowitz and A. Voronkov, eds., *Logic for Programming, Artificial Intelligence, and Reasoning, 14th International Conference, LPAR 2007*, Yerevan, Armenia, 2007, *LNAI 4790*, pp. 348–362. Springer.

32.3.6 Subterm Contextual Rewriting

Investigators: Christoph Weidenbach and Patrick Wischnewski

Contextual rewriting [1] was first implemented in the SATURATE system [2] but never matured. We have instantiated the contextual rewriting rule to subterm contextual rewriting and have implemented it in a much more sophisticated way. Our new rule is robust in the sense that invoking this rule in SPASS on the overall TPTP [3] results in an overall gain of solved problems.

The reduction rule *Subterm Contextual Rewriting* is defined as follows. We write clauses in implication notation $\Gamma \rightarrow \Delta$ denoting that the conjunction of all atoms in Γ implies the disjunction of all atoms in Δ. As usual $\prec$ is a reduction ordering total on ground terms. Let N be a clause set, $C, D \in N$, σ be a substitution then the reduction

$$\mathcal{R}\frac{D = \Gamma_1 \rightarrow \Delta_1, s \approx t \qquad C = \Gamma_2, u[s\sigma] \approx v \rightarrow \Delta_2}{\begin{array}{c}\Gamma_1 \rightarrow \Delta_1, s \approx t \\ \Gamma_2, u[t\sigma] \approx v \rightarrow \Delta_2\end{array}}$$

where the following conditions are satisfied (i) $s\sigma \succ t\sigma$, (ii) $C \succ D\sigma$, (iii) τ maps all variables from $C, D\sigma$ to fresh Skolem constants, (iv) $(\Gamma_2 \rightarrow A)\tau$ is ground subterm redundant in N for all A in $\Gamma_1\sigma$, (v) $(A \rightarrow \Delta_2)\tau$ is ground subterm redundant in N for all A in $\Delta_1\sigma$, is a subterm contextual rewriting reduction.

Our implementation of ground subterm redundancy is realized by recursively calling the reduction machinery of SPASS including a subterm contextual rewriting instance where variables of $(\Gamma_2 \rightarrow A)$, $(A \rightarrow \Delta_2)$ are implicitly treated as Skolem constants. This avoids to explicitly instantiate the side condition clauses (by τ). Further details can be found in [4, 5].

Although, the first implementation of subterm contextual rewriting lost more problems on the TPTP than it could solve, it could solve more hard problems than before. After refining the implementation with a *fault caching* as a form of (negative) dynamic programming, we could even gain twice as much problems than we lost and, additionally, could solve further hard problems. Now, there are six problems, among the problems that subterm contextual rewriting is able to solve, for which no other prover is currently able to find a proof.

References

[1] L. Bachmair and H. Ganzinger. Rewrite-based equational theorem proving with selection and simplification. *Journal of Logic and Computation*, 4(3):217–247, 1994. Revised version of Technical Report MPI-I-91-208, 1991.

[2] H. Ganzinger. The Saturate system. Available on the World-Wide Web and URL http://www.mpi-sb.mpg.de/SATURATE/Saturate.html, 1994.

[3] G. Sutcliffe and C. B. Suttner. The TPTP problem library – CNF release v1.2.1. *Journal of Automated Reasoning*, 21(2):177–203, 1998.

- [4] C. Weidenbach and P. Wischnewski. Contextual rewriting in spass. In *PAAR/ESHOL*, Sydney, Australien, 2008, *CEUR Workshop Proceedings*, vol. 373, pp. 115–124. CEUR-WS.org.
- [5] C. Weidenbach and P. Wischnewski. Contextual rewriting. Research Report MPI-I-2009-RG1-002, Max-Planck-Institut für Informatik, MPI für Informatik, Campus E 1 4, 66123 Saarbrücken, 2009.

32.4 Superposition-based Inductive Theorem Proving

Investigators: Matthias Horbach and Christoph Weidenbach

The goal of inductive theorem proving is to prove or refute formulas with respect to a distinguished model. Such problems arise naturally in many areas, such as software and hardware verification, reasoning about security protocols, or about transition systems. The higher-order status of the induction theorem makes the automation of reasoning by induction problematic in several ways: First of all, neither validity nor satisfiability is semi-decidable in higher-order logics, which means that even naive enumeration proof attempts fail. Moreover, the induction theorem cannot be handled in an efficient way comparable to the inference rules of first-order theorem provers like SPASS.

Around 1980, Lankford [6] developed the proof by consistency technique with which one can attack inductive validity without using an explicit induction scheme. Ganzinger and Stuber [2] and Comon and Nieuwenhuis [1] extended the applicability of the method to a range of specifications: Ganzinger and Stuber found a semi-decision procedure for general saturated sets of universally reductive first-order clauses and Comon and Nieuwenhuis extended their ideas to decide inductive validity in a large class of Horn theories. Both approaches use saturation as the underlying inference mechanism. Given a clause set and a conjecture, the conjecture holds in the minimal model of the clause set if its saturation terminates without generating a contradiction and without changing the minimal model at hand.

To this end, the approach of Comon and Nieuwenhuis requires the a priori computation of so called I-axiomatizations that exclude undesired models. It is not known how I-axiomatizations can be computed in general. Ganzinger and Stuber do not require I-axiomatizations – at the price of specific saturation strategies that cause their method to only terminate on a few clause classes.

We developed several superposition calculi for fixed domains that extend the scope of and partially interleave both approaches. At the heart of our methods lies a novel treatment of existentially quantified variables not by Skolemization but by constraints. While previous approaches required all variables in all clauses to be universally quantified, this enables us to incorporate existentially quantified clauses and even to allow a $\exists^*\forall^*$ quantifier alternation

in queries to be decided over a minimal model. Moreover, we cannot only reason about the minimal model but also about shared properties of all Herbrand models over a given signature [3].

In order to find terminating instances of superposition for fixed domains, we developed an extension that enables melting a diverging derivation into a finite one. For non-equational theories of universally reductive Horn clauses with linear heads over a signature of at most unary function symbols, we developed a generic algorithm that transforms a saturation-based decision procedure for the first-order validity of universally quantified query clauses into a decision procedure for inductive validity of queries with $\exists\forall^*$ quantifier alternation [5]. To our knowledge, these are the first saturation-based decidability results for inductive validity.

While saturation-based methods are ubiquitous in automated theorem proving, they are rarely used in the model building community. There, models are often represented by atomic representations or by more general disjunctions of implicit generalizations (DIGs) [7]. Based on the superposition calculus for fixed domains, we described both types of representations by saturated clause sets. This allows us not only to reprove the known decidability results using saturation, but also to extend these results to both new classes of universally quantified queries and queries containing a quantifier alternation: For DIGs, we presented a novel proof of the decidability of model equivalence and showed the decidability of validity for several classes of formulas with a $\forall^*\exists^*$ or $\exists^*\forall^*$ prefix. For atomic representations, we could even prove the decidability of validity for all such formulas [4].

References

[1] H. Comon and R. Nieuwenhuis. Induction = I-axiomatization + first-order consistency. *Information and Computation*, 159(1/2):151–186, 2000.

[2] H. Ganzinger and J. Stuber. Inductive theorem proving by consistency for first-order clauses. In J. Buchmann, H. Ganzinger, and W. J. Paul, eds., *Informatik – Festschrift zum 60. Geburtstag von Günter Hotz*, pp. 441–462. Teubner, 1992. Also in Proc. CTRS'92, LNCS 656, pp. 226–241.

- [3] M. Horbach and C. Weidenbach. Superposition for fixed domains. In *CSL*, Bertinoro, Italy, 2008, *LNCS 5213*, pp. 293–307. Springer.
- [4] M. Horbach and C. Weidenbach. Decidability results for saturation-based model building. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Computer Science. Springer. Accepted.
- [5] M. Horbach and C. Weidenbach. Deciding the inductive validity of $\forall\exists^*$ queries. Research Report MPI-I-2009-RG1-001, Max-Planck Institute for Informatics, Saarbrücken, Germany, 2009.

[6] D. S. Lankford. A simple explanation of inductionless induction. Memo mtp-14, Louisiana Technical University, Dep. of Math., Ruston, 1981.

[7] J.-L. Lassez and K. Marriott. Explicit representation of terms defined by counter examples. *Journal of Automated Reasoning*, 3(3):301–317, 1987.

32.5 Modular Deduction and Verification

Complex systems arise in a natural way in many areas such as mathematics, logic, computation, verification, databases, or artificial intelligence. Ideally, in the verification of systems which

consist of various interacting components one would like to use existing tools for verifying the components (e.g. as black-boxes), and perform the verification tasks in a modular way. Exploiting modularity in the verification of systems composed of many parallel processes allows to control, up to a certain extent, the state explosion problem. On the other hand, many problems which arise in mathematics, in the deductive verification of real-time systems or, more generally, of hybrid systems, but also in the study of complex databases, often involve reasoning in *combinations of theories*. We need efficient and accurate methods for proving validity (or, alternatively, satisfiability) of large formulae w.r.t. such complex theories. In fact, it is very important to have efficient *decision procedures*, because it is desirable (and sometimes essential) to always be sure to obtain an answer in foreseeable time. Modularity is very important in theorem proving in combinations of logical theories, since it can help in drastically reducing the number of necessary inferences.

All these aspects are addressed in our work. We analyzed possibilities of modular reasoning in combinations of theories and analyzed various applications: to deductive verification (e.g. of real time or hybrid systems), and to reasoning in terminological databases. The methods we used for this encompass:

- a systematic study of possibilities of limiting the space without losing completeness by using a goal-directed instantiation; the analysis of possibilities of increasing efficiency by using an incremental instance generation method (described in Sect. 32.5.1 and 32.5.9);
- a systematic study of possibilities of hierarchical reasoning in theory extensions, and of modular reasoning in combinations of theories (Sect . 32.5.2);
- a method for efficient generation of interpolants in certain classes of theory extensions (Sect.32.5.3);
- devising a general framework and calculus for recasting previous approaches to reasoning in theories over disjoint signatures and for extending such approaches in order to produce conflict sets (Sect. 32.5.10);
- combining SAT solvers with first order theorem provers, and combining provers for specific theories (e.g. for linear arithmetic) with first order theorem provers (Sect. 32.5.11 and 32.5.12).

Applications to efficient reasoning in various theories important in mathematical analysis, verification, non-classical logics and description logics are presented in Sect. 32.5.4, 32.5.5, 32.5.6, 32.5.7 and 32.5.8. In Section 32.5.13 we present a new topic of ongoing research in our group, on *probabilistic timed automata with first-order logic*.

We also analyzed possibilities of modular verification of complex systems: In Section 32.5.14 we report on an approach for checking properties of combinations of individual systems (such as safety) in a modular way, by only considering the individual systems, their interaction and the information that is shared between them.

32.5.1 Hierarchical Reasoning in Local Theory Extensions

Investigators: Carsten Ihlemann, Swen Jacobs, and Viorica Sofronie-Stokkermans

Many problems in computer science can be reduced to proving the satisfiability of conjunctions of literals w.r.t. a background theory. This can be a concrete theory (e.g. the theory of real

or rational numbers), the extension of a theory with additional functions (free, monotone, or recursively defined) or a combination of theories. It is therefore very important to have efficient procedures for checking the satisfiability of conjunctions of ground literals in such theories. Efficiency can be achieved for instance by:

(1) reducing the search space (preferably without losing completeness);

(2) modular reasoning, i.e., delegating some proof tasks which refer to a specific theory to provers specialized in handling formulae of that theory.

Identifying situations where the search space can be controlled without loss of completeness is of utmost importance, especially in applications where efficient algorithms (in space, but also in time) are essential. To address this problem, essentially very similar ideas occurred in various areas: proof theory (the notion of local inferences systems introduced by Givan and McAllester [3, 4]), automated deduction (the proof-theoretic research on locality and order locality by Basin and Ganzinger [1]), databases (e.g. in the study of Datalog), algebra (identifying conditions which ensure that (quasi-)varieties have a uniform word problem decidable in PTIME) and also in verification. Our results continue and considerably extend the results mentioned above, and allow us to present them from an unifying perspective.

In [2, 7] we introduced and studied a class of theory extensions, which we called *local*. In a local extension $\mathcal{T}_0 \cup \mathcal{K}$ of a base theory $\mathcal{T}_0$ with new function symbols satisfying a set $\mathcal{K}$ of axioms, when checking the satisfiability of a set of a ground clauses G w.r.t. $\mathcal{T}_0 \cup \mathcal{K}$ we do not need all instances of $\mathcal{K}$, but only the set $\mathcal{K}[G]$ consisting of those instances of $\mathcal{K}$ in which the terms starting with a new function symbol are ground terms which occurred already in $\mathcal{K}$, or occur in G. The notion of locality of a theory extension allows us to address at the same time the two aspects important for efficient reasoning mentioned above, namely restricting the search space and modular reasoning. Locality is used for restricting the search space, but as a side-effect it allows to reduce proof tasks in the extension, hierarchically, to proof tasks w.r.t. the base theory.

In [8, 9] we give an overview of results on hierarchical and modular reasoning in complex theories. We show that many theories important for computer science or mathematics fall into this class. The examples presented in [8] are theories of pointer data structures, theories of free or monotone or bounded functions. Several additional examples were studied in [5, 6, 10, 12, 13, 11, 14]. Local theories related to specific application domains will be presented in Section 32.5.4–32.5.7.

References

[1] D. Basin and H. Ganzinger. Automated complexity analysis based on ordered resolution. *J. ACM*, 48(1):70–109, 2001.

[2] H. Ganzinger, V. Sofronie-Stokkermans, and U. Waldmann. Modular proof systems for partial functions with Evans equality. *Information and Computation*, 204(10):1453–1492, 2006.

[3] R. Givan and D. McAllester. New results on local inference relations. In *Principles of Knowledge Representation and reasoning: Proceedings of the Third International Conference, KR'92*, 1992, pp. 403–412. Morgan Kaufmann Press.

[4] R. Givan and D. A. McAllester. Polynomial-time computation via local inference relations. *ACM Transactions on Computational Logic*, 3(4):521–541, 2002.

- [5] C. Ihlemann, S. Jacobs, and V. Sofronie-Stokkermans. On local reasoning in verification. In C. R. Ramakrishnan and J. Rehof, eds., *Proceedings of TACAS 2008*, Budapest, Hungary, 2008, *LNCS 4963*, pp. 265–281. Springer.
- [6] S. Jacobs and V. Sofronie-Stokkermans. Applications of hierarchical reasoning in the verification of complex systems. *Electronic Notes in Theoretical Computer Science*, 174(8):39–54, 2007.

[7] V. Sofronie-Stokkermans. Hierarchic reasoning in local theory extensions. In R. Nieuwenhuis, ed., *Automated deduction - CADE-20, 20th International Conference on Automated Deduction*, Tallinn, Estonia, 2005, *LNAI 3632*, pp. 219–234. Springer.

- [8] V. Sofronie-Stokkermans. Hierarchical and modular reasoning in complex theories: The case of local theory extensions. In B. Konev and F. Wolter, eds., *Frontiers of Combining Systems. 6th International Symposium FroCos 2007, Proceedings*, Liverpool, UK, 2007, *LNCS 4720*, pp. 47–71. Springer. Invited paper.
- [9] V. Sofronie-Stokkermans. Hierarchical and modular reasoning in complex theories: The case of local theory extensions. In S. Ranise, ed., *Proceedings of the Sixth International Workshop on First-Order Theorem Proving (FTP 2007)*, Liverpool, UK, 2007, *Technical Report, Department of Computer Science, University of Liverpool.*, vol. ULCS-07-018, p. 1. University of Liverpool. the full paper appeared in the proceedings of FroCos 2007.
- [10] V. Sofronie-Stokkermans. Efficient hierarchical reasoning about functions over numerical domains. In K. Berns and T. Breuel, eds., *KI 2008: Advances in Artificial Intelligence*, Kaiserslautern, Germany, 2008, *LNAI 5243*, pp. 135–143. Springer.

[11] V. Sofronie-Stokkermans. Locality and applications to subsumption testing and interpolation in $\mathcal{EL}$ and some of its extensions. Submitted, 2008.

- [12] V. Sofronie-Stokkermans. Locality and subsumption testing in $\mathcal{EL}$ and some of its extensions. In F. Baader, C. Lutz, and B. Motik, eds., *Proceedings of the 21st International Workshop on Description Logics (DL-2008)*, Dresden, Germany, 2008, p. 11pages. CEUR Workshop Proceedings.
- [13] V. Sofronie-Stokkermans. Locality and subsumption testing in $\mathcal{EL}$ and some of its extensions. In C. Areces and R. Goldblatt, eds., *Advances in Modal Logic, Vol.7 (Proceedings of AIML 2008)*, Nancy, France, 2008, pp. 315–339. College Publications.
- [14] V. Sofronie-Stokkermans. Locality results for certain extensions of theories with bridging functions. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer. To appear.

32.5.2 Modular Reasoning in Combinations of Local Theory Extensions

Investigator: Viorica Sofronie-Stokkermans

In applications, often it is necessary to consider complex extensions, with various types of functions (such as, for instance, extensions of the theory of real numbers with free, monotone and Lipschitz functions). It is important to have efficient methods for hierarchic and/or modular reasoning also for such combinations. Finding methods for reasoning in combinations of extensions of a base theory is far from trivial: as these are usually combinations of theories over non-disjoint signatures, classical combination results such as the Nelson-Oppen combination method [3] cannot be applied; methods for reasoning in combinations of theories over non-disjoint signatures – as studied by Ghilardi et al. [2, 1] – may also not always be applicable (unless the base theory is universal and the extensions satisfy certain model-theoretic compatibility conditions required in [2, 1]).

In [5, 6] we analyze complex theory extensions, in which various types of functions or data structures are taken into account at the same time. We investigate conditions under which locality is preserved when combining local theory extensions, and possibilities of efficient hierarchical and modular reasoning in such theory combinations.

(1) Locality of the combination makes it possible to use methods for hierarchical reasoning.

(2) By the results on interpolation in local theory extensions established in [4, 7] it follows that refuting (with respect to the combination $\mathcal{T}_1 \cup \mathcal{T}_2$ of theories sharing a common subtheory $\mathcal{T}_0$) a separated set $G_1 \wedge G_2$ of clauses can be done in a modular way if $\mathcal{T}_1 \cup \mathcal{T}_2$ is a local extension of $\mathcal{T}_0$: We can use provers for $\mathcal{T}_1$ and $\mathcal{T}_2$ as black-boxes; only ground clauses in the signature of $\mathcal{T}_0$ need to be exchanged between the provers without losing completeness. We also establish stronger conditions on $\mathcal{T}_0$ which ensure that the only information which needs to be exchanged consists of conjunctions of ground literals.

References

[1] F. Baader and S. Ghilardi. Connecting many-sorted theories. In *Proceedings of the 20th International Conference on Automated Deduction (CADE-05)*, Tallinn (Estonia), 2005, *LNAI 3632*, pp. 278–294. Springer-Verlag.

[2] S. Ghilardi. Model theoretic methods in combined constraint satisfiability. *Journal of Automated Reasoning*, 33(3-4):221–249, 2004.

[3] G. Nelson and D. C. Oppen. Simplification by cooperating decision procedures. *ACM Transations on Programming Languages and Systems*, 1(2):245–257, 1979.

[4] V. Sofronie-Stokkermans. Interpolation in local theory extensions. In U. Furbach and N. Shankar, eds., *Proceedings of IJCAR 2006*, Seattle, USA, 2006, *LNAI 4130*, pp. 235–250. Springer.

- [5] V. Sofronie-Stokkermans. Hierarchical and modular reasoning in complex theories: The case of local theory extensions. In B. Konev and F. Wolter, eds., *Frontiers of Combining Systems. 6th International Symposium FroCos 2007, Proceedings*, Liverpool, UK, 2007, *LNCS 4720*, pp. 47–71. Springer. Invited paper.

- [6] V. Sofronie-Stokkermans. Hierarchical and modular reasoning in complex theories: The case of local theory extensions. In S. Ranise, ed., *Proceedings of the Sixth International Workshop on First-Order Theorem Proving (FTP 2007)*, Liverpool, UK, 2007, *Technical Report, Department of Computer Science, University of Liverpool.*, vol. ULCS-07-018, p. 1. University of Liverpool. the full paper appeared in the proceedings of FroCos 2007.

- [7] V. Sofronie-Stokkermans. Interpolation in local theory extensions. *Logical Methods in Computer Science*, 4(4):31 pages, 2008. Special issue of LMCS dedicated to IJCAR 2006.

32.5.3 Interpolation in Local Theory Extensions

Investigator: Viorica Sofronie-Stokkermans

When reasoning in combinations of logical theories, verifying programs, or reasoning in modularly constructed databases, we want to exploit the modular structure of the systems and use mechanisms for reasoning in their "components" as black boxes. Since the components of such systems may refer to specific, partial, languages, we often need to analyze the information exchanged between the component parts in the process of deduction, and to find "local" causes

of inconsistency. In distributed databases, for instance, finding local causes of inconsistency can help in locating errors; in abstraction-based verification, finding the cause of inconsistency in a counterexample helps to rule out spurious counterexamples. Such information is usually described by *interpolants*.

It is well-known that first-order logic allows interpolation [1]. However, when computing interpolants w.r.t. a background theory, even if the theory is first-order axiomatizable the interpolant of two formulae A, B may contain an arbitrary sequence of quantifiers even if A and B are quantifier-free. It is often important to identify situations in which quantifier-free clauses have quantifier-free interpolants. In [3, 4] we study possibilities of obtaining simple interpolants in local theory extensions, i.e. quantifier-free interpolants for quantifier-free formulae. We identify situations in which it is possible to do this in a hierarchical manner, by using a prover and a procedure for generating interpolants in the base theory as "black-boxes". The method we propose in [3, 4] is general, in the sense that it can be applied to an extension $\mathcal{T}_1$ of a theory $\mathcal{T}_0$ provided that:

(i) $\mathcal{T}_0$ is convex and P-interpolating for a specified set P of predicates (cf. [3, 4]);

(ii) in $\mathcal{T}_0$ every inconsistent conjunction of ground clauses $A \wedge B$ allows a ground interpolant;

(iii) the extension clauses have a special form (for details cf. [3, 4]).

The method is *hierarchical*: the problem of finding an interpolant in $\mathcal{T}_1$ is reduced to that of finding an interpolant in the base theory $\mathcal{T}_0$. Thus, we can use the properties of $\mathcal{T}_0$ to control the form of interpolants in the extension $\mathcal{T}_1$. We identify examples of theory extensions with properties (i)–(iii), and discuss domains of applications such as:

- modular reasoning in combinations of local theories (characterization of the type of information which needs to be exchanged between provers for the individual theories for achieving completeness);
- reasoning in distributed databases (possibilities for obtaining local explanations for inconsistencies) – the applicability of these results to the study of the description logic $\mathcal{EL}$ and of its extensions is presented in more detail in [5]; and
- verification (e.g. applications to abstraction-refinement and to goal-directed over-approximation (for achieving faster termination)).

The results we obtained in [3, 4] considerably generalize existing results on interpolation in extensions of linear arithmetic with free function symbols (e.g. the results in [2]).

References

[1] W. Craig. Linear reasoning. A new form of the Herbrand-Gentzen theorem. *J. Symb. Log.*, 22(3):250–268, 1957.

[2] K. L. McMillan. An interpolating theorem prover. *Theor. Comput. Sci.*, 345(1):101–121, 2005.

[3] V. Sofronie-Stokkermans. Interpolation in local theory extensions. In U. Furbach and N. Shankar, eds., *Proceedings of IJCAR 2006*, Seattle, USA, 2006, *LNAI 4130*, pp. 235–250. Springer.

• [4] V. Sofronie-Stokkermans. Interpolation in local theory extensions. *Logical Methods in Computer Science*, 4(4):31 pages, 2008. Special issue of LMCS dedicated to IJCAR 2006.

[5] V. Sofronie-Stokkermans. Locality and applications to subsumption testing and interpolation in $\mathcal{EL}$ and some of its extensions. Submitted, 2008.

32.5.4 Hierarchical and Modular Reasoning in Theories of Data Structures

Investigators: Carsten Ihlemann, Swen Jacobs, and Viorica Sofronie-Stokkermans

In [4, 2] we present a general framework which allows to identify complex theories important in verification for which efficient reasoning methods exist. In the verification literature, locality properties were investigated in the context of reasoning in pointer data structures by McPeak and Necula [3] and in the study of fragments of the theory of arrays by Bradley, Manna and Sipma [1]. We show that locality considerations allow us to obtain parameterized decidability and complexity results for many (combinations of) theories important in verification in general and in the verification of parametric systems in particular:

- We introduce generalized notions of locality and stable locality and show that theories important in verification (e.g. the theory of arrays in [1] and the theory of pointer structures in [3]) satisfy such locality conditions.
- We present a general framework which allows to identify local theories important in verification. This allows us to also handle fragments which do not satisfy all syntactical restrictions imposed in previous papers. In particular, the axiom sets which we consider may contain alternations of quantifiers.
- We use these results to give new examples of local theories of data types.
- We explain how the decision procedures obtained this way can be used for invariant checking and bounded model checking.
- We discuss the experiments we made with an implementation.

References

[1] A. R. Bradley, Z. Manna, and H. B. Sipma. What's decidable about arrays? In E. A. Emerson and K. S. Namjoshi, eds., *Verification, Model Checking, and Abstract Interpretation, 7th International Conference (VMCAI 2006)*, 427-442, *LNCS 3855*. Springer.

• [2] C. Ihlemann, S. Jacobs, and V. Sofronie-Stokkermans. On local reasoning in verification. In C. R. Ramakrishnan and J. Rehof, eds., *Proceedings of TACAS 2008*, Budapest, Hungary, 2008, *LNCS 4963*, pp. 265–281. Springer.

[3] S. McPeak and G. C. Necula. Data structure specification via local equality axioms. In K. Etessami and S. K. Rajamani, eds., *Computer Aided Verification, 17th International Conference (CAV 2005)*, 2005, *LNCS 3576*, pp. 476–490. Springer.

• [4] V. Sofronie-Stokkermans, C. Ihlemann, and S. Jacobs. Local theory extensions, hierarchical reasoning and applications to verification. In F. Baader, B. Cook, J. Giesl, and R. Nieuwenhuis, eds., *Deduction and Decision Procedures*, Dagstuhl, Germany, 2007, *Dagstuhl Seminar Proceedings*, vol. 07401, pp. 1–22. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

32.5.5 Theories in Mathematical Analysis

Investigator: Viorica Sofronie-Stokkermans

Efficient reasoning about functions over numerical domains subject to certain properties (monotonicity, convexity, continuity or derivability) is a major challenge both in automated

reasoning and in symbolic computation. Besides its theoretical interest, it is very important for verification (especially of hybrid systems). The task of automatically reasoning in extensions of numerical domains with function symbols whose properties are expressed by first-order axioms is highly non-trivial: most existing methods are based on heuristics. Very few sound and complete methods or decidability results exist. Decidability of problems related to monotone or continuous functions over $\mathbb{R}$ were studied in [2, 3, 1]. In [5, 6] we apply methods for hierarchical reasoning we developed in [4] to the problem of checking the satisfiability of ground formulae involving functions over numerical domains.

(1) We extend the notion of locality of theory extensions in [4] to encompass additional axioms and give criteria for recognizing locality of such extensions.

(2) We give several examples, including theories of functions satisfying various monotonicity, convexity, Lipschitz, continuity or derivability conditions and certain combinations of such extensions.

(3) We illustrate the use of hierarchical reasoning to tasks such as deriving constraints between parameters which ensure (un)satisfiability, and model building (for satisfiable formulae).

We are currently using these methods in the frame of the AVACS project for the verification of hybrid systems and for the generation of Lyapunov-like criticality functions.

References

[1] D. Cantone, G. Cincotti, and G. Gallo. Decision algorithms for fragments of real analysis. I. Continuous functions with strict convexity and concavity predicates. *Journal of Symbolic Computation*, 41:763–789, 2006.

[2] H. Friedman and A. Serres. Decidability in elementary analysis, I. *Adv. in Mathematics*, 76(1):94–115, 1989.

[3] H. Friedman and A. Serres. Decidability in elementary analysis, II. *Adv. in Mathematics*, 70(2):1–17, 1990.

[4] V. Sofronie-Stokkermans. Hierarchic reasoning in local theory extensions. In R. Nieuwenhuis, ed., *Automated deduction - CADE-20, 20th International Conference on Automated Deduction*, Tallinn, Estonia, 2005, *LNAI 3632*, pp. 219–234. Springer.

- [5] V. Sofronie-Stokkermans. Efficient hierarchical reasoning about functions over numerical domains. Reports of SFB/TR 14 AVACS ATR 45, SFB/TR 14 AVACS, SFB/TR 14 AVACS, 2008. extended version of an article with the same name published in the proceedings of KI 2008.

- [6] V. Sofronie-Stokkermans. Efficient hierarchical reasoning about functions over numerical domains. In K. Berns and T. Breuel, eds., *KI 2008: Advances in Artificial Intelligence*, Kaiserslautern, Germany, 2008, *LNAI 5243*, pp. 135–143. Springer.

32.5.6 Non-classical Logics

Investigators: Carsten Ihlemann and Viorica Sofronie-Stokkermans

Another area of applications of our results is automated reasoning in non-classical logics (or their algebraic models). In [1, 2] we used methods for hierarchical reasoning in extensions of a base theory with:

- free functions, functions defined (or satisfying boundedness conditions) according to a partition of their domain of definition, described by mutually unsatisfiable formulae in the base theory;
- monotone functions (defined according to a partition of their domain of definition, or satisfying boundedness conditions);

for obtaining decision procedures and parametric complexity results for reasoning in extensions with monotone functions of general structures with a partial order (or an underlying semilattice or lattice structure) [1, 2]. Such extensions are important in non-classical logics in general and in description logics in particular. We used such procedures, in particular, for giving optimal time decision procedures for the universal theory of MV-algebras, of Gödel algebras and for other similar varieties of algebras.

References

- [1] V. Sofronie-Stokkermans and C. Ihlemann. Automated reasoning in some local extensions of ordered structures. In *Proceedings of ISMVL 2007*, Oslo, Norway, 2007, p. Article1. IEEE.
- [2] V. Sofronie-Stokkermans and C. Ihlemann. Automated reasoning in some local extensions of ordered structures. *Journal of Multiple-Valued Logic and Soft Computing*, 13(4-6):397–414, 2007.

32.5.7 TBox Subsumption in Terminological Databases

Investigator: Viorica Sofronie-Stokkermans

Description logics are logics for knowledge representation used in databases and ontologies. They provide a logical basis for modeling and reasoning about objects, classes of objects (concepts), and relationships between them (roles). Recently, tractable description logics such as $\mathcal{EL}$ [1] have attracted much interest. Although they have restricted expressivity, this expressivity is sufficient for formalizing the type of knowledge used in widely used ontologies such as the medical ontology SNOMED [11, 10]. Several papers were dedicated to studying the properties of $\mathcal{EL}$ and its extensions $\mathcal{EL}^+$ [3] and $\mathcal{EL}^{++}$ [2], and to understanding the limits of tractability in extensions of $\mathcal{EL}$. In [8, 9, 7], we proved that the subsumption problem in $\mathcal{EL}$ and $\mathcal{EL}^+$ can be expressed as a uniform word problem in certain varieties of semilattices with monotone operators (possibly satisfying certain composition laws). We identified a large class of such algebras for which the uniform word problem is decidable in PTIME. We showed that the corresponding classes of semilattices with operators have local presentations and we use methods for efficient reasoning in local theories or in local theory extensions in order to obtain PTIME decision procedures for $\mathcal{EL}$ and $\mathcal{EL}^+$. The locality considerations allowed us to present new families of PTIME logics with n-ary roles (and possibly also concrete domains) which extend $\mathcal{EL}$ and $\mathcal{EL}^+$. This allowed us to reduce, ultimately, CBox subsumption checking to checking the satisfiability of ground clauses in the theory of partially-ordered sets. We thus obtain a cubic time decision procedures for CBox subsumption in a class of extensions of $\mathcal{EL}$. In particular, we identified a PTIME extension of $\mathcal{EL}$ with two sorts, concept and num, where the concepts of sort num are interpreted as elements in the ORD-Horn, convex fragment of Allen's interval algebra [4], for which a PTIME decision procedure – based on hierarchical reasoning – can be given. Some considerations on possibilities of handling $\mathcal{EL}^{++}$ constructors and ABoxes are also presented.

The axioms which correspond, at an algebraic level, to the role inclusions in $\mathcal{EL}^+$, are exactly of the type studied in the context of hierarchical interpolation in [5, 6]. As a by-product, we thus show that the algebraic models of $\mathcal{EL}$ and $\mathcal{EL}^+$ have the ground interpolation property and infer that $\mathcal{EL}$, $\mathcal{EL}^+$, and their extensions studied in this paper have interpolation. This is especially interesting for the study of combined ontologies.

References

[1] F. Baader. Terminological cycles in a description logic with existential restrictions. In *Proc. of the 18th International Joint Conference in Artificial Intelligence*, 2003, pp. 325–330. Morgan Kaufmann.

[2] F. Baader, S. Brandt, and C. Lutz. Pushing the EL envelope. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, 2005. Morgan-Kaufmann Publishers.

[3] F. Baader, C. Lutz, and B. Suntisrivaraporn. Is tractable reasoning in extensions of the description logic EL useful in practice? *Journal of Logic, Language and Information*, 2009. To appear.

[4] B. Nebel and H.-J. Bürckert. Reasoning about temporal relations: A maximal tractable subclass of Allen's interval algebra. *Journal of the ACM*, 42(1):43–66, 1995.

[5] V. Sofronie-Stokkermans. Interpolation in local theory extensions. In U. Furbach and N. Shankar, eds., *Proceedings of IJCAR 2006*, Seattle, USA, 2006, *LNAI 4130*, pp. 235–250. Springer.

• [6] V. Sofronie-Stokkermans. Interpolation in local theory extensions. *Logical Methods in Computer Science*, 4(4):31 pages, 2008. Special issue of LMCS dedicated to IJCAR 2006.

[7] V. Sofronie-Stokkermans. Locality and applications to subsumption testing and interpolation in $\mathcal{EL}$ and some of its extensions. Submitted, 2008.

• [8] V. Sofronie-Stokkermans. Locality and subsumption testing in $\mathcal{EL}$ and some of its extensions. In F. Baader, C. Lutz, and B. Motik, eds., *Proceedings of the 21st International Workshop on Description Logics (DL-2008)*, Dresden, Germany, 2008, p. 11pages. CEUR Workshop Proceedings.

• [9] V. Sofronie-Stokkermans. Locality and subsumption testing in $\mathcal{EL}$ and some of its extensions. In C. Areces and R. Goldblatt, eds., *Advances in Modal Logic, Vol.7 (Proceedings of AIML 2008)*, Nancy, France, 2008, pp. 315–339. College Publications.

[10] K. A. Spackman. Normal forms for description logic expression of clinical concepts in SNOMED RT. *Journal of the Americal Medical Informatics Association. Symposium Supplement*, pp. 627–631, 2001.

[11] K. A. Spackman, K. E. Campbell, and R. A. Cote. SNOMED RT: A reference terminology for health care. *Journal of the Americal Medical Informatics Association. Fall Symposium Supplement*, pp. 640–644, 1997.

32.5.8 Theories of Recursively-defined Functions and of Homomorphisms and Applications to Cryptography

Investigator: Viorica Sofronie-Stokkermans

In [3], we study possibilities of reasoning about functions over theories of data types which satisfy certain recursion (or homomorphism) properties, with a focus on emphasizing possibilities of hierarchical and modular reasoning in such extensions and combinations thereof.

We start by considering theories of absolutely free data structures, and continue by studying extensions of such theories with selectors, with functions which attach scalar data to the data structures and with additional functions defined using a certain type of recursion axioms (possibly having values in a different – e.g. numeric – domain). We show that in these cases locality results can be established. This allows us to reduce the task of reasoning about the class of recursive functions we consider to reasoning in the underlying theory of absolutely free data structures (resp. in a combination of the theory of absolutely free data structures with the theory attached with the domains of the additional functions). We then show that similar results can be obtained if we relax some assumptions about the absolute freeness of the underlying theory of data types. We illustrate the ideas on a simple example from cryptography: We formalize a version of the deduction system of the Dolev and Yao protocol [2] given in [1] using a suitable chain of local theory extensions and show how hierarchical reasoning can be used for an instance of an intruder detection problem.

References

[1] H. Comon-Lundh and R. Treinen. Easy intruder deductions. In *Verification: Theory and Practice. Essays Dedicated to Zohar Manna on the Occasion of His 64th Birthday*, *LNCS 2772*, pp. 225–242. Springer, 2003.

[2] D. Dolev and A. C.-C. Yao. On the security of public key protocols. *IEEE Transactions on Information Theory*, 29(2):198–207, 1983.

- [3] V. Sofronie-Stokkermans. Locality results for certain extensions of theories with bridging functions. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer. To appear.

32.5.9 Incremental Instance Generation in Local Reasoning

Investigator: Swen Jacobs

Hierarchical reasoning in local theory extensions allows us to solve satisfiability (modulo theories) problems consisting of a ground part and sets of universally quantified axioms, called theory extensions [4]. If a given theory extension satisfies a locality property with respect to the base theory, then satisfiability of any ground problem in the extended theory is equivalent to satisfiability of the ground part together with a finite set of ground instances of the axioms in the base theory. Because the size of this set grows polynomially in the size of the ground part, eager generation of all axiom instances may not be the method of choice for large problems.

We investigated a method for incremental generation of these instances, inspired by the instantiation-based theorem proving methods by Ganzinger and Korovin [1, 2]. In these methods, instantiation of formulas is interleaved with ground satisfiability checks until either a proof or a model is found. In [3], and later in [3], we applied this idea to the framework of local theory extensions, which allows us to restrict the search for suitable instances to the finite set that is defined by the locality property. Generation of new instances is guided semantically by a partial model of the ground part, including the axiom instances generated thus far. On the one hand this should increase the probability that generated instances contribute to a proof; on the other hand it gives us an additional termination argument: if

no new instances can be generated for a given candidate model, we can already conclude that the input is satisfiable.

The incremental approach has been implemented in our tool iLoRe (see Section 32.9.5), and tested on different sets of benchmarks. We generated a set of parameterized artificial benchmarks, where the incremental approach is faster than the eager one, and increasingly so for large values of the parameter. For a set of industrial benchmarks, augmented with local axiom sets, the incremental approach is faster on the majority of problems, but there are also some for which it needs significantly more time than eager generation of all instances.

References

[1] H. Ganzinger and K. Korovin. New directions in instantiation-based theorem proving. In P. Kolaitis, ed., *18th Annual IEEE Symposium on Logic in Computer Science (LICS-03)*, Ottawa, Canada, 2003, pp. 55–64. IEEE.

[2] H. Ganzinger and K. Korovin. Theory instantiation. In M. Hermann and A. Voronkov, eds., *Logic for Programming, Artificial Intelligence, and Reasoning, 13th International Conference (LPAR'06)*, Phnom Penh, Cambodia, 2006, *LNCS 4246*, pp. 497–511. Springer.

- [3] S. Jacobs. Incremental instance generation in local reasoning. In A. Bouajjani and O. Maler, eds., *Computer Aided Verification - 21st International Conference, CAV 2009*, Grenoble, France, 2009. Springer. To appear.

[4] V. Sofronie-Stokkermans. Hierarchic reasoning in local theory extensions. In R. Nieuwenhuis, ed., *Automated deduction - CADE-20, 20th International Conference on Automated Deduction*, Tallinn, Estonia, 2005, *LNAI 3632*, pp. 219–234. Springer.

32.5.10 Combining Decision Procedures

Investigators: Duc-Khanh Tran in cooperation with Christophe Ringeissen (LORIA-INRIA Nancy), Silvio Ranise (LORIA-INRIA Nancy and University of Verona), and Hélène Kirchner (INRIA Bordeaux)

Decision procedures and constraint solvers are key components in many systems, such as automated theorem provers, expert systems, and constraint logic programming environments. Their role is to significantly augment the degree of automation of the overall system, thereby reducing user's interaction. Integrating such reasoning components requires some ingenuity as the problems tackled by such complex systems are usually (i) large, (ii) expressed over several domains, and (iii) the computed solutions may require some form of certification (e.g., for safety critical applications). In order to overcome these difficulties, in [14] we consider the following issues:

(1) the *combination* of decision procedures for signature-disjoint theories, to address (ii),

(2) the *modularity* of the computation of conflict sets or explanations of the results for such decision procedures, to address (i) and (iii).

The contributions are explained in detail below:

Combination. The research on the combination of decision procedures was started by Nelson-Oppen [9] and Shostak [12] for unions of theories with disjoint signatures. Each combination scheme makes different assumptions on the properties the theories to be combined should satisfy. The Nelson-Oppen scheme requires the theories to have a satisfiability procedures and to be such that a satisfiable formula in a component theory T is also satisfiable in an infinite model of T (*stable-infiniteness*). The Shostak combination scheme assumes that the theories admit procedures for reducing terms to canonical form (*canonizers*) and algorithms for solving equations (*solvers*). A series of papers [4, 10, 1, 7, 6, 2, 11, 3, 8] clarified the subtle issues of combining Shostak theories by studying their relationships with Nelson-Oppen theories.

In [15, 14] we provide a synthesis of Nelson-Oppen and Shostak approaches to disjoint combination by using a rule-based approach in which many recent results are recast and proved correct in a uniform, rigorous, and simple way. We formalize the combination schemata as inference systems. The applicability conditions of the inference rules are derived from the properties of the theories being combined. The lack of modularity for Shostak theories, together with the observation that the theory of equality is not a Shostak theory, but admits an efficient algorithm to derive entailed equalities suggested a possible line of investigation. We proposed the concept of *deduction completeness*. Intuitively, a deduction complete satisfiability procedure is a satisfiability procedure defined as an inference-based system with the capability of computing all the entailed elementary equalities with no overhead. We show that deduction complete inference-based satisfiability procedures can be constructed in a modular way. Another interesting feature is that they can be *built efficiently* by reusing existing techniques such as canonizers and solvers for Shostak theories and rewriting techniques. To summarize, the concept of deduction complete inference-based satisfiability procedures offers an interesting trade-off between modularity and the possibility to reuse disparate techniques to solve the satisfiability problem under a common interface.

Modular Computation of Explanations. To efficiently and correctly incorporate decision procedures into deduction systems or constraint programming environments, the capability of explaining the results of the decision procedures is crucial. For example, conflict sets (explanation of unsatisfiability) are useful to prune the search space of satisfiability modulo theories (SMT) solvers, or to direct backtracking in CLP systems, whereas explanations can be used to safely import the results of external reasoning modules (e.g., decision procedures for selected theories or unification algorithms) in skeptical proof assistants.

In [15, 14], we provide an abstract account of how to extend the Nelson-Oppen combination schema to build a satisfiability procedure capable of producing conflict sets in the union of theories T_1 and T_2, whenever the satisfiability procedures for T_1 and T_2 provide some interface capabilities. To this end, we first introduce the concept of *explanation graph*, a data structure which compactly encodes the fact that a certain equality between variables (called elementary equality) is a logical consequence of a set of elementary equalities. Explanation graphs can easily be implemented by using efficient algorithms based on the Union-Find data structure [13, 5]. Then we show how to derive *explanation engines* from satisfiability procedures that produce conflict sets in the union of the component theories.

References

[1] C. W. Barrett, D. L. Dill, and A. Stump. A generalization of Shostak's method for combining decision procedures. In A. Armando, ed., *Proc. of the 4th International Workshop on Frontiers of Combining Systems, FroCoS'02 (Santa Margherita Ligure, Italy)*, 2002, *LNCS 2309*, pp. 132–147. Springer.

[2] S. Conchon and S. Krstić. Canonization for disjoint unions of theories. In *Proc. of the 19th International Conference on Automated Deduction, CADE'03*, 2003, *LNCS 2741*, pp. 197–211. Springer.

[3] S. Conchon and S. Krstić. Strategies for combining decision procedures. In *Proc. of the 9th Int. Conference on Tools and Algorithms for the Construction and Analysis of Systems, TACAS'03*, 2003, *LNCS 2619*, pp. 537–553. Springer.

[4] D. Cyrluk, P. Lincoln, and N. Shankar. On Shostak's decision procedure for combinations of theories. In M. A. McRobbie and J. K. Slaney, eds., *Proc. of the 13th International Conference on Automated Deduction, (CADE'96), New Brunswick, NJ*, 1996, *LNAI 1104*, pp. 463–477. Springer.

[5] P. J. Downey, R. Sethi, and R. E. Tarjan. Variations on the common subexpression problem. *J. ACM*, 27(4):758–771, 1980.

[6] H. Ganzinger. Shostak light. In A. Voronkov, ed., *Proc. of the 18th International Conference on Automated Deduction, (CADE'02)*, 2002, *LNCS 2392*, pp. 332–346. Springer.

[7] D. Kapur. A rewrite rule based framework for combining decision procedures. In *Proc. of the 4th Int. Workshop on Frontiers of Combining Systems, (FroCos'02)*, 2002, *LNCS 2309*, pp. 87–102. Springer.

[8] Z. Manna and C. G. Zarba. Combining decision procedures. In *Formal Methods at the Cross Roads: From Panacea to Foundational Support*, 2003, *LNCS 2757*, pp. 381–422. Springer.

[9] G. Nelson and D. C. Oppen. Simplification by cooperating decision procedures. *ACM Transations on Programming Languages and Systems*, 1(2):245–257, 1979.

[10] H. Rueß and N. Shankar. Deconstructing Shostak. In *Proc. of the 16th Annual IEEE Symposium on Logic in Computer Science, (LICS'01), Boston, Massachusetts, USA*, 2001, pp. 19–28. IEEE Computer Society.

[11] H. Rueß and N. Shankar. Combining Shostak theories. In *Proc. of the 13th International Conference on Rewriting Techniques and Applications*, 2002, *LNCS 2378*, pp. 1–18. Springer.

[12] R. E. Shostak. A practical decision procedure for arithmetic with function symbols. *Journal of the ACM*, 26(2):351–360, 1979.

[13] R. E. Tarjan. Efficiency of a good but not linear set union algorithm. *Journal of the ACM*, 22(2):215–225, 1975.

• [14] D.-K. Tran, C. Ringeissen, S. Ranise, and H. Kirchner. Combinations of convex theories: Modularity, deduction completeness and explanation. *Journal of Symbolic Computation*, 2009. Accepted for publication.

• [15] D.-K. Tran, C. Ringeissen, S. Ranise, and H. Kirchner. Combinations of convex theories: Modularity, deduction completeness and explanation. Research Report MPI-I-2008-RG1-002, Max-Planck-Institut for Informatics, MPI-INF, Campus E 1 4, 66123 Saarbrücken, 2009.

32.5.11 Superposition Modulo Linear Arithmetic SUP(LA)

Investigators: Evgeny Kruglov and Christoph Weidenbach in cooperation with Ernst Althaus (D1)

The aim of this work is to integrate linear arithmetic into the superposition calculus for first-order logic in such a way that given a solver for linear arithmetic, whose axiomatic structure ought to remain hidden (in fact, it is not expressible in first-order logic), we are able to use it in a modular fashion, like a subroutine in a program, whereas the superposition calculus extended for the combination of the theories were applicable to the free part mostly independently of the arithmetic enrichment like the standard superposition calculus does [6, 1].

The superposition calculus can be turned into a decision procedure for several decidable first-order fragments of first-order logic, e.g. [4, 2], but linear arithmetic is not expressible in the first-order logic, therefore, the ideas of these papers cannot be used to handle the linear arithmetic case fully. There are currently three approaches known to combine superposition with linear arithmetic: SPASS+T [8] by Prevesto and Waldmann, an SMT style combination of SPASS [10] with theories, the integration of linear arithmetic in the superposition calculus suggested by Korovin and Voronkov [5], and the combination of DPLL(T) style reasoning with superposition by de Moura and Bjorner [7]. Neither approach provides completeness results of a combination. This was done by Bachmair, Ganzinger and Waldmann in their paper on hierarchic theorem proving [3] as well as for the general structure of totally ordered divisible abelian groups by Waldmann [9].

Our approach – Sup(LA) – is entirely based on the hierarchical superposition based theorem proving calculus of Bachmair, Ganzinger, and Waldmann [3], which enables the hierarchic combination of full first-order logic with a theory, which in our case is linear arithmetic. If a clause set of the combination enjoys a sufficient completeness criterion, then the calculus is even complete, meaning that a set of clauses saturated by the inference rules of the calculus is unsatisfiable if and only if it contains the empty clause.

References

- [1] E. Althaus, E. Kruglov, and C. Weidenbach. Superposition modulo linear arithmetic: SUP(LA). In S. Ghilardi and R. Sebastiani, eds., *Frontiers of Combining Systems. 7th International Symposium FroCos 2009, Proceedings*, 2009, LNCS. Springer. Accepted for publication.

[2] A. Armando, M. P. Bonacina, S. Ranise, and S. Schulz. On a rewriting approach to satisfiability procedures: Extension, combination of theories and an experimental appraisal. In B. Gramlich, ed., *Frontiers of Combining Systems, 5th International Workshop, FroCoS 2005, Vienna, Austria, September 19-21, 2005, Proceedings*, 2005, *LNCS 3717*, pp. 65–80. Springer.

[3] L. Bachmair, H. Ganzinger, and U. Waldmann. Refutational theorem proving for hierarchic first-order theories. *Applicable Algebra in Engineering, Communication and Computing (AAECC)*, 5(3/4):193–212, 1994. Earlier Version: Theorem Proving for Hierarchic First-Order Theories, in Giorgio Levi and Hélène Kirchner, editors, *Algebraic and Logic Programming, Third International Conference*, LNCS 632, pages 420–434, Volterra, Italy, September 2–4, 1992, Springer-Verlag.

[4] U. Hustadt, R. A. Schmidt, and L. Georgieva. A survey of decidable first-order fragments and description logics. *Journal of Relational Methods in Computer Science*, 1:251–276, 2004.

[5] K. Korovin and A. Voronkov. Integrating linear arithmetic into superposition calculus. In J. Duparc and T. A. Henzinger, eds., *Computer Science Logic, 21st International Workshop, CSL 2007, 16th Annual Conference of the EACSL, Lausanne, Switzerland, September 11-15, 2007, Proceedings*, 2007, *LNCS 4646*, pp. 223–237. Springer.

[6] E. Kruglov. Superposition modulo linear arithmetic. Master thesis, Universität des Saarlandes and Max-Planck-Institut für Informatik, Saarbrücken, Germany, 2008. Supervisors: E. Althaus, C. Weidenbach.

[7] L. M. de Moura and N. Bjørner. Engineering dpll(t) + saturation. In A. Armando, P. Baumgartner, and G. Dowek, eds., *Automated Reasoning, 4th International Joint Conference, IJCAR 2008*, 2008, *LNCS 5195*, pp. 475–490. Springer.

[8] V. Prevosto and U. Waldmann. SPASS+T. In G. Sutcliffe, R. Schmidt, and S. Schulz, eds., *ESCoR: FLoC'06 Workshop on Empirically Successful Computerized Reasoning*, Seattle, WA, USA, 2006, *CEUR Workshop Proceedings*, vol. 192, pp. 18–33.

[9] U. Waldmann. Superposition and chaining for totally ordered divisible abelian groups (Extended abstract). In R. Goré, A. Leitsch, and T. Nipkow, eds., *Automated reasoning, First International Joint Conference, IJCAR 2001*, Siena, Italy, 2001, *LNAI 2083*, pp. 226–241. Springer.

- [10] C. Weidenbach, R. Schmidt, T. Hillenbrand, R. Rusev, and D. Topic. System description: Spass version 3.0. In F. Pfenning, ed., *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007, *LNAI 4603*, pp. 514–520. Springer.

32.5.12 Combining SAT Solvers with First-order Theorem Provers

Investigators: Duc-Khanh Tran in cooperation with Christopher Lynch (Clarkson University, Potsdam, USA)

Deciding the satisfiability of a formula with respect to a background theory is crucial for verification. There exist specialized reasoning methods, which used to be limited to testing satisfiability of first-order formulae without quantifiers for many background theories of interest, such as lists, arrays, records, integer-offsets, and linear arithmetic, etc.; their implementations are generally known as Satisfiability Modulo Theories (SMT) solvers. Finding good heuristics for lifting SMT techniques from ground to quantified formulas is a current topic of research. For instance, [2, 6, 4] use heuristics based on the instantiation method of the theorem prover Simplify [3]. These heuristics are incomplete, i.e. they may fail to prove unsatisfiability of formulas and cannot say anything about satisfiable formulas because they instantiate universally quantified variables, and it is never known when it is safe to stop instantiating. On the other hand, there exist mature Automated Theorem Provers (ATPs) such as e.g. SPASS [10], Vampire [8], or E [9], implementing resolution and superposition calculi [1, 7] which are complete for first order logic with or without equality. A key property of such provers is that an ordering is placed on the literals which yields a drastic reduction in the search space. On several classes of problems, it causes termination and therefore the inference procedure becomes a decision procedure. However, resolution based ATPs are believed to not be as fast as SAT solvers on ground propositional problems having complex boolean structures. (The same holds for superposition based ATPs versus SMT solvers on ground equational problems.) Recently, instantiation-based ATPs have been studied – most of these provers take better advantage of the boolean structure, but not of orderings.

In [5] we study the problem of lifting SMT solvers (with equality as a background theory) from ground to quantified formulas in an efficient and complete way. We propose a novel method, that we call Satisfiability Modulo Equality with Lazy Superposition (SMELS), which combines the best of ATPs and SMT: completeness for quantified problems; and efficiency for ground problems. An ordering is used for the quantified part of the problem and an SMT solver is used for the ground part. We show how to do this without losing completeness. As far as we know, this is the first complete combination of a SAT solver with orderings. It is designed for a set of clauses that is mostly ground, with a small non-ground part representing a theory. We prove completeness of SMELS, using a nontrivial modification of Bachmair and Ganzinger's model generation technique. Completeness of SMELS ensures that one of the following will happen when applying our calculus:

(i) the original set of clauses is satisfiable, and after a finite number of steps the process will halt, giving a ground model modulo the nonground background theory; or

(ii) the original set of clauses is satisfiable, and in the limit, there is a set of clauses for which we can build a model; or

(iii) the original set of clauses is unsatisfiable, and after a finite number of steps the process will halt with an unsatisfiable set of ground clauses.

Possibilities (i) and (iii) are the most interesting compared to instantiation-based heuristics. The results of this work are published in [5].

References

[1] L. Bachmair and H. Ganzinger. Resolution theorem proving. In A. Robinson and A. Voronkov, eds., *Handbook of Automated Reasoning*, vol. 1, ch. 2, pp. 19–100. North Holland, 2001.

[2] C. Barrett and C. Tinelli. CVC3. In W. Damm and H. Hermanns, eds., *Proceedings of the 19th International Conference on Computer Aided Verification (CAV'07), Berlin, Germany*, 2007, *LNCS 4590*, pp. 298–302. Springer.

[3] D. Detlefs, G. Nelson, and J. B. Saxe. Simplify: A Theorem Prover for Program Checking. Technical Report HPL-2003-148, HP Laboratories, 2003.

[4] B. Dutertre and L. de Moura. Integrating simplex with DPLL(T). CSL Technical Report SRI-CSL-06-01, 2006.

- [5] C. Lynch and D.-K. Tran. Smels: Satisfiability modulo equality with lazy superposition. In S. S. Cha, J.-Y. Choi, K. Moonzoo, I. Lee, and M. Viswanathan, eds., *Automated Technology for Verification and Analysis 6th International Symposium, ATVA 2008*, Seoul, Korea, 2008, *LNCS 5311*, pp. 186–200. Springer.

[6] L. M. de Moura and N. Bjørner. Z3: An efficient SMT solver. In J. Rehof and C. R. Ramakrishnan, eds., *Tools and Algorithms for the Construction and Analysis of Systems, 14th International Conference, TACAS 2008*, 2008, *LNCS 4963*, pp. 337–340. Springer.

[7] R. Nieuwenhuis and A. Rubio. Paramodulation-based theorem proving. In A. Robinson and A. Voronkov, eds., *Handbook of Automated Reasoning*, pp. 371–443. Elsevier and MIT Press, 2001.

[8] A. Riazanov and A. Voronkov. The design and implementation of VAMPIRE. *AI Commun.*, 15(2):91–110, 2002.

[9] S. Schulz. E – A Brainiac Theorem Prover. *Journal of AI Communications*, 15(2/3):111–126, 2002.

[10] C. Weidenbach. SPASS version 0.49. *Journal of Automated Reasoning*, 18(2):247–252, 1997.

32.5.13 Enriched Probabilistic Timed Automata

Investigators: Arnaud Fietzke and Christoph Weidenbach in cooperation with Holger Hermanns (Saarland University)

The analysis of real-time and probabilistic computation is an active research area with applications in the verification of communication protocols and other complex systems. Probabilistic timed automata (PTA) [1] can be used to model systems with simple timing constraints and a simple, synchronization-based communication mechanism.

For more realistic models of communication, the structure of exchanged messages must be taken into account. We propose to extend PTA by adding first-order theories which describe messages and operations on them. As a general method for the verification of such enriched PTA, we consider their encoding into first-order clause logic, following [2] and [3]. We treat probabilistic and nondeterministic branching in the same way, yielding an overapproximation of reachability. The resulting clause set is then saturated using a first-order theorem prover, in our case SPASS(LA) 32.9.2. If saturation terminates, we get a finite representation of (at least) all reachable states and messages. In case we are able to extract a finite model from this representation, we can effectively construct a finite PTA that is equivalent to the original one with respect to probabilistic reachability. This automaton can then be analyzed using existing tools [1]. If no finite model can be extracted from the saturation, we try to again obtain an equivalent PTA by using abstraction techniques. If saturation does not terminate, the reachability theory is used to obtain finite reachability proofs. Each proof corresponds to one or more paths through the automaton, giving upper and lower bounds on the reachability probability. By covering all possible paths this way, we can compute exact reachability probabilities. This requires finding sufficiently many different reachability proofs, of which infinitely many may in general exist.

The main challenges are (1) finitely saturating the reachability theory, (2) finding good abstractions in case no finite model can be constructed, and (3) finding proofs corresponding to all possible paths to some state. A long-range goal is the characterization of classes of systems for which finite representations of these proofs can always be found.

Our work is part of the AVACS project, and first results were presented on the AVACS workshop 2009 in Freiburg.

References

[1] M. Kwiatkowska, G. Norman, R. Segala, and J. Sproston. Automatic verification of real-time systems with discrete probability distributions. *Theoretical Computer Science*, 282:2002, 1999.

[2] A. Nonnengart. Hybrid systems verification by location elimination. In N. Lynch and B. H. Krogh, eds., *Proceedings of the 3rd International Workshop HSCC 2000*, 2000, pp. 352–365. Springer Verlag, LNCS 1790.

[3] C. Weidenbach. Towards an automatic analysis of security protocols in first-order logic. In H. Ganzinger, ed., *Proceedings of the 16th International Conference on Automated Deduction (CADE-16)*, Trento, Italy, 1999, *LNAI 1632*, pp. 378–382. Springer.

32.5.14 Modular Verification

Investigator: Viorica Sofronie-Stokkermans

One of the problems that arise in the verification of such complex systems is that the state space may grow exponentially with the number of components. Symbolic representations of states and symbolic model checking have greatly increased the size of the systems that can be verified. However, many realistic systems are still too large to be handled. It is therefore important to find techniques that can be used to further extend the size of the systems that can be verified. One possibility is to check properties in a modular way (i.e. verify them for the individual components, infer that they also hold in the system obtained by the interconnection of the individual components, and then use them to deduce additional properties of the system). Not all properties of individual systems are preserved when interconnecting them: for instance deadlocks might occur when interconnecting deadlock free systems.

A Sheaf Theoretical Approach to the Modular Verification of Complex Systems. In [6] and later in [3, 5, 4] we address the following question, very important in verification: *Which properties of complex systems can be checked in a modular way?* To answer this question, we use an analogy with phenomena in topology and algebraic geometry, where "locally defined" objects are studied, which need to be "patched" to "globally defined" objects. An important tool used in algebraic geometry in the study of such phenomena is sheaf theory. We show that also in the study of complex systems sheaf theory allows us to establish links between "local" and "global" description of systems as well as between "local" and "global" properties of systems. (In our setting, the open sets model the component collections of subsystems; the topological space we use expresses how the interacting systems share the information; the *data* are descriptions of the *states*, *possible parallel actions*, *transitions*, or of the *behavior* of the systems, or of the local time of the systems.) We show that, given a family of interacting systems, states, actions, transitions, behavior in time have the property that any collection of compatible data (i.e. "local" states, actions, transitions, resp. behavior) assigned to a family of interacting systems can be "glued together" to a "global" element (state, action, transition, or behavior). In other words, we show that states, actions, transitions, behavior in time can often be modeled by sheaves over this topological space which describes how the interacting systems share the information. We consider possibilities of modeling behavior either by traces or by partially commutative monoids. Many properties of systems can be expressed as assertions about states, actions, transitions, behavior in time. The sheaf semantics allows us to prove that those properties of systems that can be expressed by *cartesian axioms* are preserved after interconnecting the systems (cf. also [6]).

Applications to the Verification of the Safety of a System of Trains. In [3, 5, 4], all ideas are illustrated on a running example involving a family of interacting controllers controlling a subsets of consecutive trains on a linear, loop-free rail track.

By applying the theoretical results we show that – for that particular description – we can decompose the initial family of trains into a 'covering' family $\{S_i \mid i \in \{2, \ldots, n\}\}$ of systems consisting of two successor trains each (cf. Fig. 32.2). Here S_i is the system consisting of train $i-1$ and train i on the track (with all the corresponding control rules); the intersection of S_i and S_{i+1} consists of the one-train system formed by train i. Let U consist of this family of systems together with their intersections. The colimit of this family is the initial system S of trains we considered. By our results, if collision freeness can be guaranteed for all the systems in U, then the system S is collision free. For suitably chosen minSpeed, maxSpeed and update interval Δt it can be proved that the 2-train systems are collision free (for an automatic proof ideas from [1, 2] can be used).

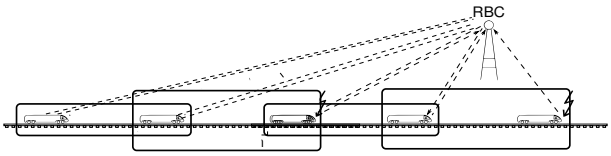


Figure 32.2: System of trains on linear track: decomposition in two-element subsystems

We plan to use similar ideas for proving safety of geographically distributed systems by considering a covering of the space with regions controlled by compatible controllers (Fig. 32.3).

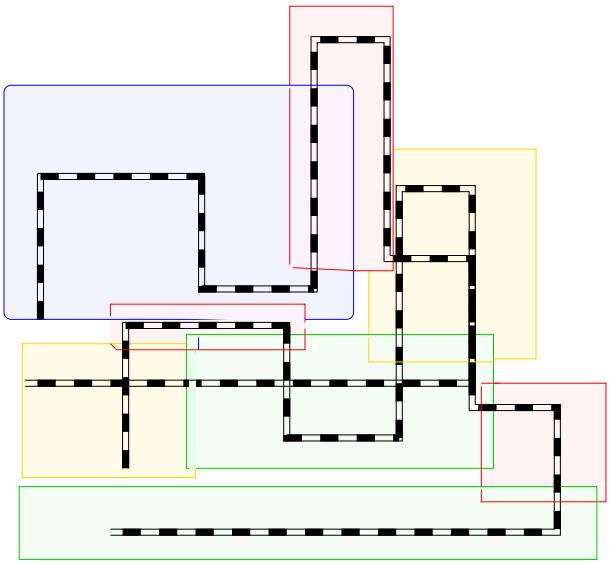


Figure 32.3: Geographically distributed train controllers

First steps in this directions are currently for the verification of a system of trains on a complex line track (cf. Section 32.8.3). We plan to further investigate the link between the "locality" aspects described here and the locality results in automated reasoning presented in Section 32.5.1.

References

[1] S. Jacobs and V. Sofronie-Stokkermans. Applications of hierarchical reasoning in the verification of complex systems. In B. Cook and R. Sebastiani, eds., *PDPAR'06: Pragmatical Aspects of Decision Procedures in Automated Reasoning*, Seattle, USA, 2006, pp. 15–26.

- [2] S. Jacobs and V. Sofronie-Stokkermans. Applications of hierarchical reasoning in the verification of complex systems. *Electronic Notes in Theoretical Computer Science*, 174(8):39–54, 2007.

[3] V. Sofronie-Stokkermans. Sheaves and geometric logic in concurrency. In *Proceedings of the Eighth Workshop on Geometric and Topological Methods in Concurrency (GETCO 2006)*, Bonn, Germany, 2006. -.

- [4] V. Sofronie-Stokkermans. Sheaves and geometric logic and applications to modular verification of complex systems. Reports of SFB/TR 14 AVACS ATR 46, SFB/TR 14 AVACS, SFB/TR 14 AVACS, 2008.

- [5] V. Sofronie-Stokkermans. Sheaves and geometric logic and applications to modular verification of complex systems. *Electronic Notes in Theoretical Computer Science*, 230:161–187, 2009.

[6] V. Sofronie-Stokkermans and K. Stokkermans. Modeling interaction by sheaves and geometric logic. In G. Ciobanu and G. Paun, eds., *Proceedings of the 12th International Symposium Fundamentals of Computation Theory (FCT-99)*, Iasi, Romania, 1999, *LNCS 1684*, pp. 512–523. Springer.

32.6 Decision Procedures

A useful measure for the potential of a calculus, methodology for the automation of logic is its potential to actually create decision procedures. The first two sections (Section 32.6.1, 32.6.2) reflect our results presented in Section 32.5 and Section 32.4, respectively, from this perspective whereas we also present new results in unification theory (Section 32.6.3).

32.6.1 Decision Procedures based on Hierarchical Reasoning

Investigators: Carsten Ihlemann, Swen Jacobs, and Viorica Sofronie-Stokkermans

We used the possibilities for hierarchical reasoning in local theory extensions described in Sections 32.5.1–32.5.5 for devising decision procedures for the universal fragment (and also for slightly more general fragments) of the following theories:

- Theories of functions satisfying monotonicity conditions and (guarded) boundedness conditions over partially ordered domains with a decidable universal theory. The details are presented in Sect. 32.5.1 and 32.5.6.
- The universal theory of MV-algebras and Gödel algebras, which are algebraic counterparts of many many-valued logics having as set of truth values the interval $[0, 1]$ (e.g. Łukasiewicz logics and Gödel logics). The details are presented in Sect. 32.5.6.
- The framework of local theory extensions allows us to generate (or recognize) in a systematic way a large class of local theory extensions related to data structures (including various theories of pointers and arrays) and give decision procedures for testing satisfiability of ground formulae in such extensions. The details are presented in Sect. 32.5.4.

- Various theories of functions from mathematical analysis (monotone or convex/concave functions; functions satisfying Lipschitz and bi-Lipschitz conditions and possibly also continuity and derivability conditions). The details are presented in Sect. 32.5.5.
- Certain theories of recursive functions and homomorphisms. The details are presented in Sect. 32.5.8.
- We give PTIME decision procedures (with optimal complexity) for the TBox subsumption in description logics $\mathcal{EL}, \mathcal{EL}^+$ and some extensions thereof with n-ary roles and concrete domains. The details are presented in Sect. 32.5.7.

The decision procedures are based on the fact that in Ψ-local extensions of a base theory $\mathcal{T}_0$ (where Ψ is a closure operator for ground terms with the property that if T is a finite set of ground terms then $\Psi(T)$ is finite) we can reduce the task of checking satisfiability of a set G of ground clauses to the task of checking satisfiability of a formula G' in the base theory $\mathcal{T}_0$. Then satisfiability w.r.t. $\mathcal{T}_1$ is decidable for all sets of ground clauses G for which the corresponding formula G' in the base signature is finite and belongs to a fragment $\mathcal{F}_0$ of $\mathcal{T}_0$ for which checking satisfiability is decidable. This hierarchical reduction also allows us to give parameterized complexity results for the theory extension: Assume that for every set of ground clauses G, the formula G' in the base signature obtained after the hierarchical reduction is finite and belongs to a fragment $\mathcal{F}_0$ of $\mathcal{T}_0$ for which checking satisfiability is decidable. Assume that for any formula in $\mathcal{F}_0$ of size m, its satisfiability can be checked in time $g(m)$. Then the complexity of checking satisfiability of a formula G w.r.t. $\mathcal{T}_1$ is of order $g(n^k)$, where $n = |\Psi_{\mathcal{K}}(G)|$ and k is the maximum among the number of extension terms for each clause in $\mathcal{K}$ (but at least 2).

32.6.2 Decision Procedures for Inductive Queries

Investigators: Matthias Horbach and Christoph Weidenbach

We explored saturation-based approaches to inductive theorem proving in the tradition of Ganzinger and Stuberand Comon and Nieuwenhuis.For a class of Horn clauses, we developed a generic algorithm that transforms a saturation-based decision procedure for the first-order validity of universally quantified clauses into a decision procedure for inductive validity of queries with $\exists\forall^*$ quantifier alternation.

We also introduced saturation-based methods to extend known decidability results in model building. For models represented by disjunctions of implicit generalizations, we presented a novel proof of the decidability of model equivalence and showed the decidability of validity for several classes of formulas with a $\forall^*\exists^*$ or $\exists^*\forall^*$ prefix. For atomic representations, we could even prove the decidability of the validity for all such formulas.

A more thorough overview of the results can be found in section 32.4.

32.6.3 Decision Procedures for Unification based on Natural Dualities

Investigator: Viorica Sofronie-Stokkermans

Unification is important in computer science; it is used e.g. in resolution-based theorem proving and in term rewriting to deal with certain equational axioms, such as associativity and commutativity; or in knowledge representation. The unification problem has thoroughly been

studied for equationally defined theories characterized by axioms such as associativity, commutativity, distributivity, associativity-commutativity, associativity-commutativity-idempotency; and for several theories related to algebra (abelian groups, commutative and Boolean rings, semilattices, Boolean algebras, primal algebras, discriminator varieties). For details cf. [1] and the bibliography cited there. It can be quite complicated to use equational reasoning for deciding unification. Also direct checking by using descriptions of the free algebras is often not feasible, even for relatively simple finitely generated varieties, due to the complexity of the free algebras.

In [5] we consider the problem of unification in certain finitely-generated classes of algebras of the form $\mathcal{V} = ISP(\underline{P})$, where $\underline{P}$ is a finite algebra, for which natural dualities exist. Assume $\mathcal{V} = ISP(\underline{P})$. The idea of *natural duality theorems* [3] is to seek an "alter ego" for the generating algebra $\underline{P}$, which we require to be a topological structure $\underline{\underline{P}} = (P, \tau, R)$ on the underlying set P of $\underline{P}$, so chosen that there is a dual equivalence between $\mathcal{V}$ and a suitable category Sp of topological relational structures of the same type as $\underline{\underline{P}}$. The goal is to obtain a concrete representation of any $\mathbf{A} \in \mathcal{V}$ as the algebra of continuous R-preserving maps from $D(\mathbf{A}) = \mathsf{Hom}_{\mathcal{V}}(\mathbf{A}, \underline{P})$. The existence of such representation theorems allows us to concretely describe the free algebras in $\mathcal{V}$ and their duals and to reduce the unification problems to the problem of checking the satisfiability of a system of finite domain constraints, where the finite domain is, essentially, a description of the relational space $\underline{\underline{P}}$.

A similar class of clauses (MV-clauses) were studied in the context of many-valued logics in [2, 4]. We extend the results in [4] to the type of satisfiability problems we obtain using the reduction described above. We obtain a specially tuned resolution calculus which can be used to give decision procedures for the positive theory of the class of algebras we consider.

Alternatively, we can use the fact that the constraints are over a known, finite domain and reduce the initial unification problem to SAT checking.

References

[1] F. Baader and W. Snyder. Unification theory. In A. Robinson and A. Voronkov, eds., *Handbook of Automated Reasoning*, vol. I, ch. 8, pp. 445–532. Elsevier Science, 2001.

[2] M. Baaz and C. G. Fermüller. Resolution-based theorem proving for manyvalued logics. *J. Symbolic Computation*, 19(4):353–391, 1995.

[3] D. Clark and B. Davey. *Natural dualities for the working algebraist.* Cambridge studies in adv. math., Vol.57. Cambridge Univ. Press, 1998.

[4] H. Ganzinger and V. Sofronie-Stokkermans. Chaining techniques for automated theorem proving in many-valued logics. In *Proceedings of the 30th IEEE International Symposium on Multiple-Valued Logic (ISMVL-00)*, Portland, Oregon, 2000, pp. 337–344. IEEE.

• [5] V. Sofronie-Stokkermans. On unification in certain finitely generated varieties of algebras. In E. Contejean, ed., *Proceedings of the 21th International Workshop on Unification (UNIF 2007)*, Paris, France, 2007, pp. 1–5.

32.7 First-order Model Checking

The analysis of hybrid systems faces the difficulty of having to address not only the continuous dynamics of mechanical, electrical and other physical phenomena, but also the intricacies of

discrete switching. Both of these two constituents of hybrid systems alone often pose a major challenge for verification approaches, and their combination is of course by no means simpler. For instance, the behavior of a car or airplane is usually beyond the scope of mathematically precise assessment, even if attention is restricted to only one particular aspect like the functioning of a braking assistant. Even though the continuous behavior might in such a case be rather simple – at least after it has been simplified by introducing worst-case assumptions to focus on the safety-critical aspects –, through the interaction with discrete-state control the result is in most cases unmanageable by present-day techniques.

32.7.1 FOMC with Large Discrete State Spaces

Investigators: Swen Jacobs and Uwe Waldmann in cooperation with Werner Damm (Oldenburg University), Christoph Scholl (Freiburg University), et al.

In (Damm et al. [1, 2]) we address the analysis of hybrid systems with large discrete state spaces. Such systems arise naturally in application classes where the overall control of system dynamics rests with a finite-state supervisory control, and states represent knowledge about the global system status. In our approach, we profit from the independence of the supervisory control and the continuous sections, using adequate techniques for each of the two constituents in a hybrid procedure. We do so by representing discrete states *symbolically*, as in symbolic model checking, and combine this with a first-order logic representation of the continuous part. In that way, unnecessary distinctions between discrete states can be avoided and efficiency gained. Sets of discrete states are represented in an efficient format for boolean functions, in our case functionally reduced AND-Inverter graphs (AIGs) [3]. The state vectors are extended by additional components referring to linear first-order constraints, resulting in linear constraint AIGs (LinAIGs). Model checking works essentially as in [4] on the discrete part, while in parallel for the continuous part a Hoare-like calculus is applied. Our framework has been applied both to discrete-time model checking [2] and to continuous-time model checking [1], using Loos-Weispfenning quantifier elimination for the latter. Test vector generation for fast inequality checks of boolean combinations of constraints, implication checks for linear constraints, and advanced redundancy elimination techniques are used both in the context of subsumption checks and for keeping state set representations as compact as possible.

References

- [1] W. Damm, S. Disch, H. Hungar, S. Jacobs, J. Pang, F. Pigorsch, C. Scholl, U. Waldmann, and B. Wirtz. Exact state set representations in the verification of linear hybrid systems with large discrete state space. In K. S. Namjoshi, T. Yoneda, T. Higashino, and Y. Okamura, eds., *Automated Technology for Verification and Analysis, 5th International Symposium, ATVA 2007*, Tokyo, Japan, 2007, *LNCS 4762*, pp. 425–440. Springer.

- [2] W. Damm, S. Disch, H. Hungar, J. Pang, F. Pigorsch, C. Scholl, U. Waldmann, and B. Wirtz. Automatic verification of hybrid systems with large discrete state space. In S. Graf and W. Zhang, eds., *Automated Technology for Verification and Analysis, 4th International Symposium, ATVA 2006*, Beijing, China, 2006, *LNCS 4218*, pp. 276–291. Springer.

[3] A. Mishchenko, S. Chatterjee, R. Jiang, and R. K. Brayton. FRAIGs: A unifying representation for logic synthesis and verification. Technical report, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 2005.

[4] F. Pigorsch, C. Scholl, and S. Disch. Advanced unbounded model checking by using AIGs, BDD sweeping and quantifier scheduling. In *9th ITG/GI/GMM Workshop "Methoden und Beschreibungssprachen zur Modellierung und Verifikation von Schaltungen und Systemen"*, 2006.

32.7.2 Combining Zonotopes and AIGs for FOMC

Investigators: Willem Hagemann and Uwe Waldmann in cooperation with Werner Damm (Oldenburg University), Christoph Scholl (Freiburg University), et al.

Zonotopes have been proposed by Girard et al. [2, 3] as a data structure for efficient first-order model checking for time invariant systems described by linear differential equations, in particular in the presence of bounded inputs or disturbances. We are currently working on integrating zonotope-based flowpipe computation into the LinAIG framework for first-order model checking [1]. Rather than storing entire flowpipes in AIGs, we use a switching-plane-based approach, computing all intersections of flowpipes with potential entry hyperplanes for the respective mode. When the discrete part passes control to the continuous part, a conversion from intersections of linear constraints into zonotopes becomes necessary (and vice versa). For that purpose we decompose convex polytopes into lower dimensional linearly independent factors. For each factor we compute a zonotope (or a union of zonotopes) using an extensible collection of exact and over-approximating methods. In the two dimensional case we use an exact method to represent a triangle as a union of three zonotopes. Since every polygon can be disassembled into a union of triangles this method results in an exact representation of polygons. Additionally there is an efficient method to determine the smallest enclosing zonotope due to Guibas [4]. For higher dimension the exact methods are subjected to the vertex explosion thus we currently limit ourself to determine the smallest enclosing parallelotope which can be efficiently converted to a zonotope.

References

- [1] W. Damm, S. Disch, H. Hungar, S. Jacobs, J. Pang, F. Pigorsch, C. Scholl, U. Waldmann, and B. Wirtz. Exact state set representations in the verification of linear hybrid systems with large discrete state space. In K. S. Namjoshi, T. Yoneda, T. Higashino, and Y. Okamura, eds., *Automated Technology for Verification and Analysis, 5th International Symposium, ATVA 2007*, Tokyo, Japan, 2007, *LNCS 4762*, pp. 425–440. Springer.

[2] A. Girard. Reachability of uncertain linear systems using zonotopes. In *Proceedings of the 8th International Workshop on Hybrid Systems: Computation and Control*, 2005, pp. 542–556.

[3] A. Girard, C. Le Guernic, and O. Maler. Efficient computation of reachable sets of linear time-invariant systems with inputs. In *Proceedings of the 9th International Workshop on Hybrid Systems: Computation and Control*, 2006, pp. 257–271.

[4] L. J. Guibas, A. Nguyen, and L. Zhang. Zonotopes as bounding volumes. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, 2003, pp. 803–812.

32.8 Applications

We have started to put more emphasis on applying our results in practice. Both to show impact and to get ideas for further useful development directions for our theoretical work. The successful application of our results and tools ranges from detailed, (almost) full fledged real-world scenarios (Section 32.8.1–32.8.2) to more abstract, conceptually important problems (Section 32.8.3-32.8.6).

32.8.1 Feature Modelling

Investigators: Christoph Weidenbach and Patrick Wischnewski in cooperation with Christian Dressler and Georg Rock (PROSTEP AG)

Variant management is currently a "hot topic" in industry. A typical application are intelligent BOMs (bills of materials) where in addition to the decomposition structure of a BOM, typically represented by a tree, side conditions can be expressed orthogonal to the tree structure. An example are intelligent BOMs for car construction as they are needed to specify all configurations of a car or even the overall build. Configuration services are today offered via the Internet by almost any car manufacturer. A typical side condition may then be that the sports model has to come with a powerful gasoline engine and cannot come with 5 doors. A second typical example are intelligent software BOMs where different software modules are composed to eventually fit a specific product out of a product line. Side conditions then concern for example compatibility or interface properties between the modules.

Feature models [1] are a standard way to represent intelligent BOMs. We have designed a variant of SPASS that is able to decide satisfiability and consequence properties for the translation of feature models into propositional and first-order logic. The tool is integrated into the feature modelling tool VEIA of the PROSTEP AG and currently evaluated for industrial usage. Compared to actual tools available on the market it turned out by experiments done by PROSTEP that SPASS can handle satisfiability for almost twice as large problems as the best current industrial tools. These tools eventually rely on SAT-solving, BDD-solving, or constraint-solving technology. The SPASS core inference engine is more than a factor 1000 slower than a state-of-the-art SAT solver on propositional problems. It's our simplification machinery that is able to successfully reduce formulas resulting from feature models to a large extend.

References

[1] K. Czarnecki, S. Helsen, and U. W. Eisenecker. Formalizing cardinality-based feature models and their specialization. *Software Process: Improvement and Practice*, 10(1):7–29, 2005.

32.8.2 Analysis of Authorizations in SAP R/3

Investigators: Manuel Lamotte and Christoph Weidenbach

Today many companies use an ERP (Enterprise Resource Planning) system such as the SAP R/3 system to run their daily business ranging from financial issues down to the actual control

of a production line. These systems are very complex from the view of administration and authorization already due to their sheer size. Hence they include a high potential for errors.

The idea of our effort is to automatically analyze the authorization concept of the SAP R/3 system. To this end we construct a corresponding model in first-order logic. This model can be used to check the existence of errors automatically, e.g. a contradiction between given authorizations and a valid business regulation. For these checks we employed the theorem prover SPASS which has been developed at the Max Planck Institute for Informatics. We used the purchase process as an example to explore the model construction which is a typical constituent of the SAP R/3 system.

Our experiments on the purchase process show that (dis)proofs of errors in the SAP R/3 authorization concept can be done fully automatically with SPASS. In case of an error the prover shows the involved parts leading to the error. In case of the absence of an error it presents a model. The model enables to prove arbitrary ground queries with SPASS. Further details can be found in [1, 2].

Future tasks are the proof of decidability of the model in general as well as the integration of linear arithmetic expressions (which can for example be used to represent monetary amounts) using SPASS(LA) (Sect. 32.5.11 and 32.9.2).

References

- [1] M. Lamotte. Analysis of Authorizations in SAP R/3. Masters thesis, Fachhochschule Trier, 2008.
- [2] M. Lamotte-Schubert and C. Weidenbach. Analysis of Authorizations in SAP R/3. In *Workshop on First-Order Theorem Proving (FTP)*, 2009. Accepted for publication.

32.8.3 Verification of a Family of ETCS Case Studies

Investigators: Swen Jacobs and Viorica Sofronie-Stokkermans in cooperation with Johannes Faber (University Oldenburg)

We continued our work on applications of local reasoning in automatic verification. Some of the simplest tasks in the verification of reactive, real time and hybrid systems is automatic invariant checking, i.e. checking whether a given formula is an inductive invariant. In our work, we formally represented the states of the systems to be verified using abstract data structures. This allowed us to pass in an elegant way from verification of several finite instances of problems (modeled by finite-state systems) to general verification results, in which sets of states are represented using formulae in first-order logic. We showed that the updates of the state of the system as well the safety properties define, in many cases, local theory extensions, Therefore the problems which occur in invariant checking and bounded model checking can often be expressed as the problem of checking the satisfiability of sets of ground clauses w.r.t. a theory which can be represented in a natural way as a chain of local theory extensions. Therefore – both for invariant and for bounded model checking – we could use results in hierarchic theorem proving. By combining the possibilities of hierarchical reasoning with quantifier elimination and techniques from symbolic computation we obtained "reverse engineering" results for this special type of parametric systems, i.e. we generated relations between the parameters of these systems which guarantee safety.

Our Running Example

In the context of the AVACS project, our running example is taken from the European train control standard (ETCS). We study a system of trains on a rail track, controlled by a so-called radio block center (RBC). The RBC is responsible for a given part of the track and communicates with all trains that are within this area (Fig 32.4). Trains may enter and leave

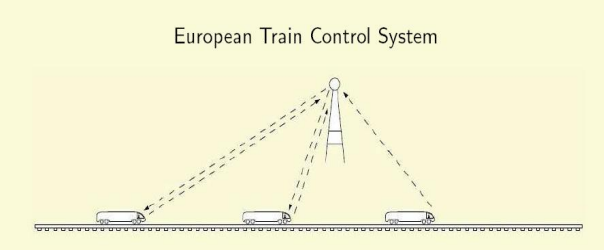


Figure 32.4: ETCS case study: Trains on a linear rail track controlled by RBC

the area, given that a certain maximum number of trains on the track is not exceeded. Every train reports its position to the RBC in given time intervals and the RBC communicates to every train how far it can safely move, based on the position of the preceding train. It is then the responsibility of the trains to adjust their speed between given minimum and maximum speeds. The requirement is to prove that, with correctly chosen minimum and maximum speeds as well as small enough time intervals for communication, one can guarantee that no collisions between trains happen. In our first treatment of this case study [5, 6], local reasoning allowed us to prove safety of such a system without restricting the number of allowed trains, but we had to keep it simple in other dimensions of complexity: we abstracted away from the details of communication, we only considered the normal mode of operation (without emergencies), and our track was a single line without junctions. For storing the train positions we used sorted arrays (or, more generally, functions $f : \mathbb{Z} \to \mathbb{R}$); safety was expressed as strict monotonicity of these functions. In [8] we illustrate the applicability of the verification tasks for an axiomatization of safety using a stronger version of strict monotonicity, which takes the lengths of the trains into account and ensures that the distance between two consecutive trains is large enough, given the length of the trains.

Extensions of the Case Study. We extended the system model [1] by taking into account emergencies and emergency messages. In this work, the initial model was given in the combined specification language CSP-OZ-DC (COD) [2], which allows to define systems with respect to their control flow (Communicating Sequential Processes, CSP), their data changes (Object-Z, OZ) and their timing aspects (Duration Calculus, DC). Previous work on verification of COD specifications [3, 7] did not incorporate data structures like lists or arrays, and parameters were not allowed in the DC part, i.e. the timing aspects of the system had to be defined with respect to fixed values. We showed that systems specified in COD, with complex data types and timing parameters, can be translated to transition constraint systems, which can then in turn be checked for safety using locality properties of the logical theories involved.

Currently, we are developing another extension of this case study, where we want to use our locality results for pointer data structures [4] to allow for a more detailed model of both

the rail track and the trains on it. Pointer data structures enable us to model not only a single track, but a whole network of tracks, consisting of track segments which may have different properties and can be connected in different ways. An example of a track network is represented in Fig. 32.5. A sequence of trains on a route through this network can now be

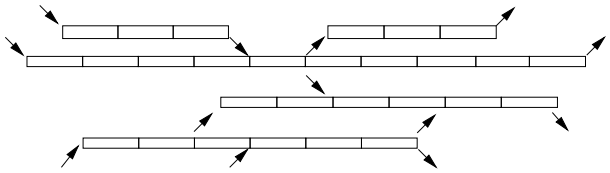


Figure 32.5: Example of track network

modeled as a linked list of train objects, allowing us to remove trains from everywhere in the list, whereas before we could only allow leaving or entering trains at the ends of the array. In particular, we can now model junctions of tracks and allow for different behavior of trains on different track segments.

Using similar methods as before, we want to prove safety of an arbitrary route through such a track network, where trains may enter or leave the route at junctions. Another goal is to have a completely automatic tool-chain which takes such a COD specification and the corresponding safety requirements, translates the specification to a transition constraint system and proves safety using H-PILoT, our prover for local theory extensions (see Section 32.9.6).

References

- [1] J. Faber, S. Jacobs, and V. Sofronie-Stokkermans. Verifying CSP-OZ-DC specifications with complex data types and timing parameters. In J. Davies and J. Gibbons, eds., *Proceedings of IFM 2007: Integrated Formal Methods*, Oxford, UK, 2007, *LNCS 4591*, pp. 233–252. Springer.

[2] J. Hoenicke. *Combination of Processes, Data, and Time*. PhD thesis, University of Oldenburg, Germany, 2006.

[3] J. Hoenicke and P. Maier. Model-checking of specifications integrating processes, data and time. In J. S. Fitzgerald, I. J. Hayes, and A. Tarlecki, eds., *FM 2005*, 2005, *LNCS 3582*. Springer.

- [4] C. Ihlemann, S. Jacobs, and V. Sofronie-Stokkermans. On local reasoning in verification. In C. R. Ramakrishnan and J. Rehof, eds., *Proceedings of TACAS 2008*, Budapest, Hungary, 2008, *LNCS 4963*, pp. 265–281. Springer.

[5] S. Jacobs and V. Sofronie-Stokkermans. Applications of hierarchical reasoning in the verification of complex systems. In B. Cook and R. Sebastiani, eds., *PDPAR'06: Pragmatical Aspects of Decision Procedures in Automated Reasoning*, Seattle, USA, 2006, pp. 15–26.

- [6] S. Jacobs and V. Sofronie-Stokkermans. Applications of hierarchical reasoning in the verification of complex systems. *Electronic Notes in Theoretical Computer Science*, 174(8):39–54, 2007.

[7] R. Meyer, J. Faber, and A. Rybalchenko. Model checking duration calculus: A practical approach. In *ICTAC*, 2006, *LNCS 4281*, pp. 332–346. Springer.

- [8] V. Sofronie-Stokkermans, C. Ihlemann, and S. Jacobs. Local theory extensions, hierarchical reasoning and applications to verification. In F. Baader, B. Cook, J. Giesl, and R. Nieuwenhuis, eds., *Deduction and Decision Procedures*, Dagstuhl, Germany, 2007, *Dagstuhl Seminar Proceedings*,

vol. 07401, pp. 1–22. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

32.8.4 Parametric Verification of a Process Controlling Protocol

Investigators: Carsten Ihlemann, Swen Jacobs, and Viorica Sofronie-Stokkermans

In [2] we considered the following types of verification problems. Consider a parametric number m of processes. The priorities associated with the processes (non-negative real numbers) are stored in an array p. The states of the processes – enabled (1) or disabled (0) are stored in an array a. At each step only the process with maximal priority is enabled, its priority is set to x and the priorities of the waiting processes are increased by y (see also Fig. 32.6).

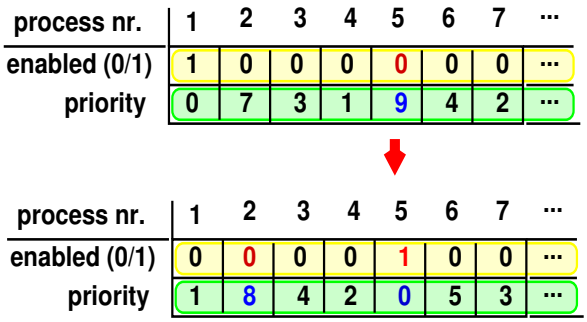


Figure 32.6: A process controlling protocol

This can be expressed using a set $\mathsf{Update}(a, p, a', p')$ of axioms which may contain alternations of quantifiers (because of the necessity of specifying maxima). We considered a version of the problem in which x and y are considered to be parameters. We may need to check whether if at the beginning the priority list is injective, i.e. formula $(\mathsf{Inj})(p)$ holds:

$\mathsf{Inj}(p) \quad \forall i, j(1 \leq i \leq m \wedge 1 \leq j \leq m \wedge i \neq j \rightarrow p(i) \neq p(j))$

then it remains injective after the update, i.e. check the satisfiability of:

$(\mathbb{Z} \cup \mathbb{R}_+ \cup \{0, 1\}) \wedge \mathsf{Inj}(p) \wedge \mathsf{Update}(a, p, a', p') \wedge 1 \leq c \leq m \wedge 1 \leq d \leq m \wedge c \neq d \wedge p'(c) = p'(d).$

We may need to check satisfiability of the formula under certain assumptions on the values of x and y (for instance if $x = 0$ and $y = 1$), or to determine constraints on x and y for which the formula is (un)satisfiable. The problem above is a satisfiability problem for a formula with (alternations of) quantifiers in a combination of theories. SMT provers heuristically compute ground instances of the problems, and return unsatisfiable if a contradiction is found, and unknown if no contradiction can be derived from these instances. It is important to find a set of ground instances which are sufficient for deriving a contradiction if one exists. The formula above does not belong to fragments of the theory of arrays previously known to be decidable such as the one in [1]: $\mathsf{Inj}(p)$ contains the premise $i \neq j$; $\mathsf{Update}(a, p, a', p')$ contains $\forall\exists$ axioms. Our idea for handling this verification task (and similar ones) [2] was to represent the involved theories as chains of theory extensions: Let $\mathcal{T}_0$ be the many-sorted combination of

the theory of integers (for indices), of real numbers (priorities), and $\{0, 1\}$ (enabled/disabled). We consider:

(i) The extension $\mathcal{T}_1$ of $\mathcal{T}_0$ with the functions $a : \mathbb{Z} \to \{0, 1\}$ (a free function) and $p : \mathbb{Z} \to \mathbb{R}_+$ satisfying $\mathsf{Inj}(p)$;

(ii) The extension $\mathcal{T}_2$ of $\mathcal{T}_1$ with the functions $a' : \mathbb{Z} \to \{0, 1\}$, $p' : \mathbb{Z} \to \mathbb{R}_+$ satisfying the update axioms $\mathsf{Update}(a, p, a', p')$.

We show that both extensions have a locality property which allows us to use determined instances of the axioms without loss of completeness; the satisfiability problem w.r.t. $\mathcal{T}_2$ can be hierarchically reduced to a satisfiability problem w.r.t. $\mathcal{T}_1$ and then to a satisfiability problem w.r.t. $\mathcal{T}_0$.

Several other case studies were analyzed in a similar way (including the verification of several programs manipulating data structures, e.g. algorithms for insertion in sorted lists or arrays [2], or sorting algorithms). The benchmarks are presented and discussed in [3]

References

[1] A. R. Bradley, Z. Manna, and H. B. Sipma. What's decidable about arrays? In E. A. Emerson and K. S. Namjoshi, eds., *Verification, Model Checking, and Abstract Interpretation, 7th International Conference (VMCAI 2006)*, 427-442, *LNCS 3855*. Springer.

- [2] C. Ihlemann, S. Jacobs, and V. Sofronie-Stokkermans. On local reasoning in verification. In C. R. Ramakrishnan and J. Rehof, eds., *Proceedings of TACAS 2008*, Budapest, Hungary, 2008, *LNCS 4963*, pp. 265–281. Springer.

- [3] C. Ihlemann and V. Sofronie-Stokkermans. System description: H-PILoT version 1.5. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer.

32.8.5 Local Reasoning and Abstraction-refinement

Investigator: Viorica Sofronie-Stokkermans in cooperation with Andrey Rybalchenko (MPI-SWS)

The algorithm for hierarchical computation of interpolants described in [3, 4] can be used in abstraction-refinement based verification. In many applications from verification we need to describe the change of certain variables parametrically, using functions with certain properties (e.g. monotonicity and boundedness). If using classical abstraction methods from verification, spuriousness of counterexamples obtained after abstraction can not be detected if we regard (as many systems do) these functions as free function symbols. An example, referring to a parametrically described controller for a water tank, is described in [3, 4]. Therefore, it was necessary to develop methods for efficiently generating ground interpolants for extensions with functions satisfying various types of axioms. This greatly widens the area of application of abstraction-based verification methods to classes of parametric real time and hybrid systems in which the evolution of certain (continuous) variables is characterized parametrically, only by means of axioms. The algorithm for hierarchical computation of interpolants proposed in [3, 4] was applied (and tuned) to the special problem of efficiently computing interpolants with a simple form for extensions of linear arithmetic with free function symbols in joint work of

with Andrey Rybalchenko [2]. This implementation is integrated into the predicate discovery procedure of the software verification tools Blast (http://embedded.eecs.berkeley.edu/blast/) and ARMC (http://www.mpi-sb.mpg.de/∼rybal/armc). Several experiments with Blast on Windows device drivers provide a direct comparison with the existing tool FOCI [1], and show promising running times in favor of our constraint based approach.

Besides the application to verification by abstraction-refinement, computation of Craig interpolants has other potential applications (e.g. to goal-directed over-approximation for achieving faster termination, or to automatic invariant generation) in the verification of such types of parametric systems which we plan to explore in future work.

References

[1] K. L. McMillan. An interpolating theorem prover. *Theor. Comput. Sci.*, 345(1):101–121, 2005.

- [2] A. Rybalchenko and V. Sofronie-Stokkermans. Constraint solving for interpolation. In B. Cook and A. Podelski, eds., *Verification, Model Checking and Abstract Interpretation, 8th International Conference, VMCAI 2007*, Nice, France, 2007, *LNCS 4349*, pp. 346–362. Springer.

[3] V. Sofronie-Stokkermans. Interpolation in local theory extensions. In U. Furbach and N. Shankar, eds., *Proceedings of IJCAR 2006*, Seattle, USA, 2006, *LNAI 4130*, pp. 235–250. Springer.

- [4] V. Sofronie-Stokkermans. Interpolation in local theory extensions. *Logical Methods in Computer Science*, 4(4):31 pages, 2008. Special issue of LMCS dedicated to IJCAR 2006.

32.8.6 Hierarchical Reasoning for Description Logics

Investigator: Viorica Sofronie-Stokkermans

The results on hierarchical reasoning in local theory extensions turned out to also be helpful in identifying tractable description logics. In [6, 7, 5] we show that subsumption problems in lightweight description logics (such as $\mathcal{EL}$ and $\mathcal{EL}^+$) can be expressed as uniform word problems in classes of semilattices with monotone operators. We use possibilities of efficient local reasoning in such classes of algebras, to obtain uniform PTIME decision procedures for CBox subsumption in $\mathcal{EL}$, $\mathcal{EL}^+$ and extensions thereof. These locality considerations allow us to present a new family of (possibly many-sorted) logics which extend $\mathcal{EL}$ and $\mathcal{EL}^+$ with n-ary roles and/or numerical domains. The results in [3, 4] allowed us to show that the algebraic models of $\mathcal{EL}$ and $\mathcal{EL}^+$ have ground interpolation; this property is important when considering combinations of ontologies which use the language of $\mathcal{EL}$ or $\mathcal{EL}^+$. In future work we would like to test the implementation of the method for hierarchical reasoning [1] and also the implementation using lazy instantiation [2] on existing $\mathcal{EL}$ or $\mathcal{EL}^+$ ontologies for terminological reasoning in medicine such as e.g. SNOMED [9, 8].

References

- [1] C. Ihlemann and V. Sofronie-Stokkermans. System description: H-PILoT version 1.5. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer.

- [2] S. Jacobs. Incremental instance generation in local reasoning. In A. Bouajjani and O. Maler, eds., *Computer Aided Verification - 21st International Conference, CAV 2009*, Grenoble, France, 2009. Springer. To appear.

[3] V. Sofronie-Stokkermans. Interpolation in local theory extensions. In U. Furbach and N. Shankar, eds., *Proceedings of IJCAR 2006*, Seattle, USA, 2006, *LNAI 4130*, pp. 235–250. Springer.

- [4] V. Sofronie-Stokkermans. Interpolation in local theory extensions. *Logical Methods in Computer Science*, 4(4):31 pages, 2008. Special issue of LMCS dedicated to IJCAR 2006.

[5] V. Sofronie-Stokkermans. Locality and applications to subsumption testing and interpolation in $\mathcal{EL}$ and some of its extensions. Submitted, 2008.

- [6] V. Sofronie-Stokkermans. Locality and subsumption testing in $\mathcal{EL}$ and some of its extensions. In F. Baader, C. Lutz, and B. Motik, eds., *Proceedings of the 21st International Workshop on Description Logics (DL-2008)*, Dresden, Germany, 2008, p. 11pages. CEUR Workshop Proceedings.

- [7] V. Sofronie-Stokkermans. Locality and subsumption testing in $\mathcal{EL}$ and some of its extensions. In C. Areces and R. Goldblatt, eds., *Advances in Modal Logic, Vol.7 (Proceedings of AIML 2008)*, Nancy, France, 2008, pp. 315–339. College Publications.

[8] K. A. Spackman. Normal forms for description logic expression of clinical concepts in SNOMED RT. *Journal of the Americal Medical Informatics Association. Symposium Supplement*, pp. 627–631, 2001.

[9] K. A. Spackman, K. E. Campbell, and R. A. Cote. SNOMED RT: A reference terminology for health care. *Journal of the Americal Medical Informatics Association. Fall Symposium Supplement*, pp. 640–644, 1997.

32.9 Software

Our tools can be divided into two categories. Firstly, the systems SPASS (Section 32.9.1), SPASS(T) (Section 32.9.2), SPASS+T (Section 32.9.3), and WALDMEISTER (Section 32.9.4) are build around the superposition (completion) calculus for first-order (equational) logic. Whereas SPASS and WALDMEISTER concentrate on pure first-order (equational) logic, SPASS(T) and SPASS+T provide two different solutions for reasoning in a combination of first-order logic with some theory T. Secondly, the systems iLoRe (Section 32.9.5) and H-PILoT (Section 32.9.6) implement SMT and hierarchically based theory combination approaches, respectively.

32.9.1 SPASS

Investigators: Arnaud Fietzke, Christoph Weidenbach, Patrick Wischnewski, Dilyana Dimova, Thomas Hillenbrand, Rohit Kumar, Rostislav Rusev, and Dalibor Topic in cooperation with Renate Schmidt (Univ. Manchester)

SPASS is an automated theorem prover for full first-order logic with equality and a number of non-classical logics. We describe an overview of the developments for SPASS 3.0 as well as our recent developments in SPASS 3.5.

New in SPASS 3.0 [5] are facilities for supporting automated reasoning in a large class of related logics which we refer to as *EML* logics (extended modal logics). These include (traditional) propositional modal logics such as $K_{(m)}$, $KD_{(m)}$, $KT4_{(m)}$ etc., which are widely used for studying and formalizing e.g. multi-agent systems, but have many applications in other areas of computer science as well as mathematics, linguistics and philosophy. *EML* logics also include dynamic modal logics which are PDL-like modal logics in which the modal operators are parameterized by relational formulas. These can be used to formalize dynamic

notions such as actions or programs and are useful in linguistic and AI applications. Examples of dynamic modal logics are Boolean modal logic, tense logic, information logics, logics expressing inaccessibility and sufficiency as well as a large class of description logics. The *EML* class further includes relational logics, i.e. logical versions of Tarski's relation algebras. SPASS handles these logics by translation to first-order logic.

One important change coming with the recent version 3.5 [4] is the change from a GPL to a BSD style license. We have been asked by several companies for this change and now eventually did it. Actually, it took some effort as it required the (re)implementation of several SPASS modules as we were so far relying on some code distributed under a GPL license, for example the command line parser up to version 3.0. Starting from SPASS version 3.5 it will be distributed under a "non-virulent" BSD style license.

The most important enhancements based on advances in theory are the integration of subterm contextual rewriting (Section 32.3.6) and improved split backtracking (Section 32.3.1). Further enhancements are:

Faster Parsing: Until SPASS version 3.0 the overall input machinery was done for "small" input files. Due to our and the general interest in large files, e.g., expressing real-world finite domain theories, we have reimplemented the overall SPASS parsing technology. For example, we now can parse and build the CNF for an 1 MB input file like SEU410+2 from TPTP version 3.5.0 with *full* FLOTTER CNF translation and reduction in about 30 seconds.

TPTP Input Syntax Support: Starting from SPASS version 3.5 we support TPTP input files via the new flag `-TPTP` [2]. As TPTP input files may contain include commands, they are resolved by looking in the local directory or the value of the `TPTP` environment variable.

Include Commands: SPASS input files may now also contain include directives. They are resolved at parse time and include files are looked up in the local directory as well as in the directory bound to the `SPASSINPUT` environment variable.

tptp2dfg: The new tool tptp2dfg translates input files from TPTP into SPASS syntax. Includes can either be expanded or translated into SPASS include directives controlled by the flag `-include`.

Sort Module: In SPASS sorts are used for soft typing [1, 3]. We reimplemented the module such that it is now about 10-times faster and extended its scope. For example, the new module reduces a clause $S(x), T(x), \Gamma \to \Delta$ where x does not occur in Γ, Δ to $\Gamma \to \Delta$ in the presence of the three clauses $\to S(a)$, $S(x) \to S(f(x))$, and $\to T(f(a))$.

Symmetric Reduction: Until SPASS version 3.0 some of the more sophisticated rewrite reductions were only implemented in the forward direction. We now added also the backward direction for all reduction rules.

References

[1] H. Ganzinger, C. Meyer, and C. Weidenbach. Soft typing for ordered resolution. In *Proceedings of the 14th International Conference on Automated Deduction, CADE-14*, Townsville, Australia, 1997, *LNAI 1249*, pp. 321–335. Springer.

[2] G. Sutcliffe and C. B. Suttner. The TPTP problem library – CNF release v1.2.1. *Journal of Automated Reasoning*, 21(2):177–203, 1998.

[3] C. Weidenbach. Combining superposition, sorts and splitting. In A. Robinson and A. Voronkov, eds., *Handbook of Automated Reasoning*, vol. 2, ch. 27, pp. 1965–2012. Elsevier, 2001.

- [4] C. Weidenbach, D. Dimova, A. Fietzke, M. Suda, and P. Wischnewski. Spass version 3.5. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Computer Science. Springer. Accepted.

- [5] C. Weidenbach, R. Schmidt, T. Hillenbrand, R. Rusev, and D. Topic. System description: Spass version 3.0. In F. Pfenning, ed., *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007, *LNAI 4603*, pp. 514–520. Springer.

32.9.2 SPASS(T)

Investigators: Evgeny Kruglov and Christoph Weidenbach

SPASS(T) is the hierarchic extension of SPASS by a theory T [1]. The advantage of a hierarchic approach is an abstract completeness result for the calculus that mainly depends on a sufficient completeness property. Basically, all function terms that are not solely terms of the theory T have to be reducible to a pure T term.

We have intensively studied the case where T is the theory of linear arithmetic. For linear arithmetic we developed appropriate redundancy criteria instantiating the abstract calculus. For an implementation of the combination with linear arithmetic we coupled SPASS with linear programming solvers such as QSopt(`http://www2.isye.gatech.edu/~wcook/qsopt/index.html`) as well as SMT solvers such as Z3 [3]. The prototype combination SPASS(LA) is available from the SPASS homepage (`http://www.spass-prover.org/prototypes/index.html`). Our own experiments showed the potential of the approach, e.g., we could automatically prove inductive safety properties of hybrid systems. Further research will try to mature this specific combination.

We also did first experiments studying the case where T is a non-linear arithmetic theory using the HySAT (`http://hysat.informatik.uni-oldenburg.de/`) system [2]. It turned out that the structure of the problems delegated to HySAT does not perfectly match its reasoning capabilities. This is a topic of current work. in the AVACS project.

References

[1] L. Bachmair, H. Ganzinger, and U. Waldmann. Refutational theorem proving for hierarchic first-order theories. *Applicable Algebra in Engineering, Communication and Computing (AAECC)*, 5(3/4):193–212, 1994. Earlier Version: Theorem Proving for Hierarchic First-Order Theories, in Giorgio Levi and Hélène Kirchner, editors, *Algebraic and Logic Programming, Third International Conference*, LNCS 632, pages 420–434, Volterra, Italy, September 2–4, 1992, Springer-Verlag.

[2] M. Fränzle and C. Herde. HySAT: An efficient proof engine for bounded model checking of hybrid systems. *Formal Methods in System Design*, 30(3):179–198, 2007.

[3] L. M. de Moura and N. Bjørner. Z3: An efficient SMT solver. In J. Rehof and C. R. Ramakrishnan, eds., *Tools and Algorithms for the Construction and Analysis of Systems, 14th International Conference, TACAS 2008*, 2008, *LNCS 4963*, pp. 337–340. Springer.

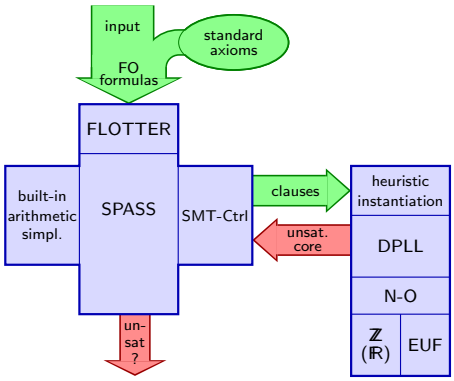


Figure 32.7: SPASS+T architecture

32.9.3 SPASS+T

Investigators: Uwe Waldmann and Stephan Zimmer

Standard first-order theorem provers are notoriously bad at dealing with integer or real arithmetic – encoding numbers in binary or unary is not really a viable solution in most application contexts. SPASS+T (Prevosto and Waldmann [2]) adds arithmetic reasoning capabilities to SPASS in three complementary ways: First, it uses a set of standard axioms such as $\forall x.\,(x + 0 = x)$ or $\forall x, y.\,(x - y) + y = x)$. Second, built-in simplification rules are employed to reduce numeric subexpressions and to find solutions for variables, so that, for instance, a clause $\forall x.\,(\neg\, x + 2 = 5 + 1 \vee p(x))$ is replaced by $\forall x.\,(\neg\, x = 4 \vee p(x))$ and then by $p(4)$. Third, SPASS is linked to an arbitrary SMT (*Satisfiability Modulo Theories*) procedure for arithmetic and free function symbols, such as Yices (Dutertre and De Moura [1]). In this combination, SPASS uses its deduction rules to generate formulas as usual; in addition, the SMT procedure is repeatedly called with the sets of formulas generated so far. As soon as one of the two systems encounters a contradiction, the problem is solved. The latest version of SPASS+T features support for splitting inferences, a built-in ordered chaining inference rule, a new term ordering, exact integer and rational arithmetic, and proof documentation for SMT subproofs [3].

References

[1] B. Dutertre and L. de Moura. A fast linear-arithmetic solver for DPLL(T). In T. Ball and R. B. Jones, eds., *Computer Aided Verification, 18th International Conference, CAV 2006*, Seattle, WA, USA, 2006, *LNCS 4144*, pp. 81–94. Springer-Verlag.

[2] V. Prevosto and U. Waldmann. SPASS+T. In G. Sutcliffe, R. Schmidt, and S. Schulz, eds., *ESCoR: FLoC'06 Workshop on Empirically Successful Computerized Reasoning*, Seattle, WA, USA, 2006, *CEUR Workshop Proceedings*, vol. 192, pp. 18–33.

- [3] S. Zimmer. Intelligent combination of a first order theorem prover and smt procedures. Diploma thesis, Universität des Saarlandes, 2007.

32.9.4 WALDMEISTER

Investigator: Thomas Hillenbrand

> *I did invest many profitable hours in using the equational theorem prover,* WALDMEISTER*: it is small, yet very effective on many problems involving equational deductions.* – Dana S. Scott [2, p. VIII]

For automated reasoning in structures that are axiomatized by equations, Knuth-Bendix completion and its refinements are the method of choice today. The WALDMEISTER system is among the strongest implementations thereof, as demonstrated repeatedly in last years' competitions at the Conference on Automated Deduction. Furthermore, it is now also part of the MATHEMATICA computer algebra system of Wolfram Research, Inc. On this occasion, invited talks have been given in 2008 at the Conference on Rewriting Techniques and Applications in Linz and at the Workshop on Empirically Successful Automated Reasoning for Mathematics in Birmingham. Recent system developments include interfacing with the TPTP tools for proof processing. On the application side, automated reasoning tools have an increasing impact on the theory of loops and quasigroups [1], where WALDMEISTER assists in obtaining new results.

References

[1] J. D. Philipps and D. Stanovský. Automated theorem proving in loop theory. In G. Sutcliffe, S. Colton, and S. Schulz, eds., *Proceedings of the Workshop on Empirically Successful Automated Reasoning for Mathematics*, 2008, *CEUR Workshop Proceedings*, vol. 378, pp. 42–54.

[2] F. Wiedijk, ed. *The Seventeen Provers of the World*, *LNAI 3600*. Springer-Verlag, 2006.

32.9.5 iLoRe

Investigator: Swen Jacobs

Satisfiability problems in local extensions of theories can be reduced from first-order problems to ground problems by finite instantiation of quantified variables [3]. The ground satisfiability problems can then be decided by existing SMT solvers. However, the number of needed axiom instances grows polynomially with the ground part of the original problem. Thus, generating all of them and solving a problem with thousands of axiom instances may not be the method of choice for large problems.

iLoRe implements an incremental approach for generating these instances [1]. Instead of generating all instances before checking satisfiability, instance generation is interleaved with satisfiability checks of the ground part, including ground axiom instances generated thus far. This ground satisfiability check is done by a blackbox SMT solver (Z3 [2] at the moment, but others can be integrated). If the ground part of the problem is still satisfiable, we retrieve a model from the SMT solver and use it to guide generation of new axiom instances. The prover terminates when either no new instances can be generated for the given model, in which case the original problem is satisfiable, or when the ground part becomes unsatisfiable, in which case the original problem was also unsatisfiable.

Experiments show that for a large class of examples, incremental generation of instances is more efficient than generating all instances up front.

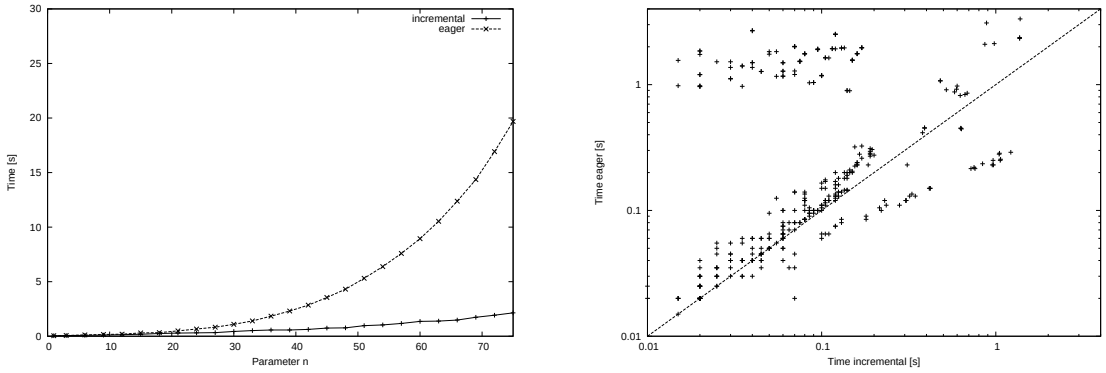


Figure 32.8: Runtime comparison of the eager and incremental approaches to local reasoning on two different sets of benchmarks

References

- [1] S. Jacobs. Incremental instance generation in local reasoning. In A. Bouajjani and O. Maler, eds., *Computer Aided Verification - 21st International Conference, CAV 2009*, Grenoble, France, 2009. Springer. To appear.

[2] L. M. de Moura and N. Bjørner. Z3: An efficient SMT solver. In J. Rehof and C. R. Ramakrishnan, eds., *Tools and Algorithms for the Construction and Analysis of Systems, 14th International Conference, TACAS 2008*, 2008, *LNCS 4963*, pp. 337–340. Springer.

[3] V. Sofronie-Stokkermans. Hierarchic reasoning in local theory extensions. In R. Nieuwenhuis, ed., *Automated deduction - CADE-20, 20th International Conference on Automated Deduction*, Tallinn, Estonia, 2005, *LNAI 3632*, pp. 219–234. Springer.

32.9.6 H-PILoT

Investigators: Carsten Ihlemann, Viorica Sofronie-Stokkermans

H-PILoT (Hierarchical Proving by Instantiation in Local Theory extensions) implements hierarchical reasoning in extensions of logical theories with a set of clauses containing additional function symbols [2]: deduction problems in the theory extension are reduced to deduction problems in the base theory [3, 4, 5, 1]. This reduction can always be carried out in polynomial time (w.r.t. the number of input clauses).

A general prover such as SPASS, a standard SMT solver or a specialized prover is then called by H-PILoT for testing the satisfiability of the formulae obtained after the reduction. The hierarchical reduction used in H-PILoT is always sound; it is complete for the class of so-called local extensions of a base theory.

H-PILoT provides a decision procedure for testing satisfiability of ground formulas if the clauses obtained by this reduction belong to a decidable fragment of the base theory. If the result of the reduction is a satisfiable problem, H-PILoT can also be used for model generation.

Many real-life problems in verification can be decided by H-PILoT in this manner because the problem setting can be seen as a local extension over a base theory known to be decidable.

Common local extensions include the theory of monotone/bounded functions, the theory of pointers and the theory of arrays with integer indices. For the latter two H-PILoT also automatically checks whether necessary syntactic criteria are met to keep track of whether H-PILoT is a full decision procedure for the respective fragment or only decides unsatisfiable problems.

Experiments confirm the practical usefulness of H-PILoT for verification purposes. In particular in the face of satisfiable problems over local theory extensions, there H-PILoT always provides the right answer. In such types of extensions, H-PILoT is a decision procedure whereas completeness of other SMT-solvers is not guaranteed.

References

- [1] C. Ihlemann, S. Jacobs, and V. Sofronie-Stokkermans. On local reasoning in verification. In C. R. Ramakrishnan and J. Rehof, eds., *Proceedings of TACAS 2008*, Budapest, Hungary, 2008, *LNCS 4963*, pp. 265–281. Springer.
- [2] C. Ihlemann and V. Sofronie-Stokkermans. System description: H-PILoT version 1.5. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer.

[3] V. Sofronie-Stokkermans. Hierarchic reasoning in local theory extensions. In R. Nieuwenhuis, ed., *Automated deduction - CADE-20, 20th International Conference on Automated Deduction*, Tallinn, Estonia, 2005, *LNAI 3632*, pp. 219–234. Springer.

[4] V. Sofronie-Stokkermans. Interpolation in local theory extensions. In U. Furbach and N. Shankar, eds., *Proceedings of IJCAR 2006*, Seattle, USA, 2006, *LNAI 4130*, pp. 235–250. Springer.

- [5] V. Sofronie-Stokkermans. Hierarchical and modular reasoning in complex theories: The case of local theory extensions. In B. Konev and F. Wolter, eds., *Frontiers of Combining Systems. 6th International Symposium FroCos 2007, Proceedings*, Liverpool, UK, 2007, *LNCS 4720*, pp. 47–71. Springer. Invited paper.

32.10 Academic Activities

32.10.1 Journal Positions

Viorica Sofronie-Stokkermans:

– *Special issue of the Journal of Symbolic Computation* (Guest Editor)

32.10.2 Conference and Workshop Positions

Membership in Program Committees

Thomas Hillenbrand:

– *Workshop on Empirically Successful Automated Reasoning for Mathematics, ESARM*, Birmingham, United Kingdom, July/August 2008,

– *Workshop on Practical Aspects of Automated Reasoning, PAAR-2008*, Sydney, Australia, August 2008.

– *7th International Workshop on the Implementation of Logics, IWIL*, Doha, Qatar, November 2008,

Viorica Sofronie-Stokkermans:

– *Automated Deduction: Decidability, Complexity, Tractability, ADDCT'07*, Bremen, Germany, July 2007 (Program Chair).
– *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods, TABLEAUX 2007*, Aix en Provence, France, July 2007,
– *Deduktionstreffen 2008*, Saarbrücken, Germany, March 2008,
– *4th International Joint Conference on Automated Reasoning, IJCAR 2008*, Sydney, Australia, August 2008,
– *Complexity, Expressibility, and Decidability in Automated Reasoning, CEDAR'08*, Sydney, Australia, August 2008 (Program Chair),
– *International Workshop on First-Order Theorem Proving, FTP 2009*, Oslo, Norway, July 2009 (Program Chair)
– *18th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods, TABLEAUX 2009*, Oslo, Norway, July 2009,
– *22nd International Workshop on Description Logics, DL2009*, Oxford, United Kingdom, July 2009,
– *Automated Deduction: Decidability, Complexity, Tractability, ADDCT'09*, Montreal, Canada, August 2009 (Program Chair),
– *International Symposium on Frontiers of Combining Systems, FroCoS 2009*, Trento, Italy, September 2009,

Uwe Waldmann:

– *Workshop on Empirically Successful Automated Reasoning in Large Theories, ESARLT*, Bremen, Germany, July 2007.
– *Workshop on Empirically Successful Automated Reasoning for Mathematics, ESARM*, Birmingham, United Kingdom, July/August 2008,
– *Workshop on Practical Aspects of Automated Reasoning, PAAR-2008*, Sydney, Australia, August 2008,

Christoph Weidenbach:

– *Tenth International Conference on Foundations of Software Science and Computation Structures, FoSSaCS 2007*, Braga, Portugal, March/April 2007,
– *14th International Conference on Logic for Programming Artificial Intelligence and Reasoning , LPAR 2007*, Yerevan, Armenia, October 2007,
– *International Workshop on First-Order Theorem Proving, FTP 2009*, Oslo, Norway, July 2009.
– *22nd International Conference on Automated Deduction, CADE 22*, Montreal, Canada, August 2009,

Membership in Organizing Committees

Viorica Sofronie-Stokkermans:

- *Automated Deduction: Decidability, Complexity, Tractability, ADDCT'07*, Bremen, Germany, July 2007,
- *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods, TABLEAUX 2007*, Aix en Provence, France, July 2007 (Workshop Chair).
- *Symbolic Computation and Deduction in System Design and Verification, Special session at ACA 2008*, Hagenberg, Austria, July 2008,
- *Complexity, Expressibility, and Decidability in Automated Reasoning, CEDAR'08*, Sydney, Australia, August 2008,
- *Automated Deduction: Decidability, Complexity, Tractability, ADDCT'09*, Montreal, Canada, August 2009,

Christoph Weidenbach:

- *Summer School 2008: Verification Technology, Systems & Applications*, Max Planck Institute for Informatics, Saarbrücken, September 2008.
- *AVACS Winter School 2009*, University of Freiburg, November 2008.

32.10.3 Invited Talks and Tutorials

Thomas Hillenbrand:

- *Fast Equational Reasoning with Waldmeister*, Invited talk, International Conference on Rewriting Techniques and Applications (RTA), Linz, July 2008.
- *Reasoning in Equational Logic with Waldmeister*, Invited talk, Workshop on Empirically Successful Automated Reasoning for Mathematics (ESARM), Birmingham, July 2008.

Viorica Sofronie-Stokkermans:

- *Hierarchical and modular reasoning in complex theories*, Invited talk, Università degli Studi di Verona, May 2007.
- *Modularität in der Verifikation komplexer Systeme*, Invited talk, Universität Gießen, June 2007.
- *Hierarchical and modular reasoning in complex theories: The case of local theory extensions*, Invited talk, FroCos'07 and FTP'07 (joint session), Liverpool, September 2007.
- *Modularität im Automatischen Beweisen und in der Verifikation komplexer Systeme*, Invited talk, Universität Paderborn, September 2007.
- *Modularity in automated reasoning and the verification of complex systems*, Invited talk, University of Manchester, October 2007.
- *Reasoning in complex theories*, Tutorial, KI 2008, Kaiserslautern, September 2008.

- *Hierarchical and modular reasoning in complex theories*, Invited talk, Instituto Superior Technico, Lisbon, March 2009.
- *Hierarchical and modular reasoning in complex theories*, Tutorial, 22nd International Conference on Automated Deduction, Montreal, Canada, August 2009.

Uwe Waldmann:

- *SPASS+T*, Invited talk, EPFL Lausanne, December 2007.

Christoph Weidenbach:

- *Automatic Analysis of LAN Infrastructures*, Invited talk, ENS Cachan, Paris, June 2007.
- *Theorem Proving for Feature Models*, Invited Talk, PROSTEP AG, Darmstadt, March 2008.
- *Automated Reasoning*, Invited Lecture, Tsinghua University, Peking, May 2008.

32.10.4 Other Academic Activities

Viorica Sofronie-Stokkermans:

- *Reviewer for the German Science Foundation (DFG).*

Christoph Weidenbach:

- *Member of the Steering Committee for the French-German Computer Science Cooperation Agreement between INRIA, CNRS, University of Metz, University of Nancy 1, University of Nancy 2, Institut National Polytechnique de Lorraine at Nancy, University of Saarbrücken, University of Kaiserslautern, Fraunhofer Institute for Experimental Software Engineering (IESE) Kaiserslautern, Max Planck Institute for Informatics, Max Planck Institute for Software Systems, DFKI.*
- *Member of the Selection Committee of the Saarbrücken Graduate School in Computer Science.*
- *Member of the CASC ATP System Competition Panel 2009.*

32.11 Teaching Activities

Winter Semester 2007/2008

Courses:
Unix for Advanced Users (U. Waldmann)

Proseminars:
Decision Procedures based on SAT (C. Weidenbach, M. Lamotte)

Summer Semester 2008

Courses:

Automated Reasoning (C. Weidenbach)

Seminars:

Decision Procedures for Logical Theories (V. Sofronie-Stokkermans, U. Waldmann)

Winter Semester 2008/2009

Courses:

Advanced C Programming (C. Weidenbach together with S. Hack from Saarland University)

Summer Semester 2009

Courses:

Reasoning in complex theories and applications (V. Sofronie-Stokkermans, Advanced Course at ESSLLI 2009)

Selected topics in Automated Reasoning (V. Sofronie-Stokkermans)

Theses

Dilyana Dimova: Propositional Abduction, Bachelor thesis, 2007 (C. Weidenbach)

Christian Dressler: First-Order Proof Documentation, Bachelor thesis, 2007 (C. Weidenbach)

Arnaud Fietzke: Labelled Splitting, Master's thesis, 2007 (C. Weidenbach)

Patrick Wischnewski: Contextual Rewriting in SPASS, Master's thesis, 2007 (C. Weidenbach)

Stephan Zimmer: Intelligent Combination of a First Order Theorem Prover and SMT Procedures, Diploma thesis, 2007 (U. Waldmann, C. Weidenbach)

Rostislav Rusev: Bitvector Reasoning with SPASS, Master's thesis, 2008 (T. Hillenbrand, C. Weidenbach)

Manuel Lamotte: Analysis of Authorizations in SAP R/3, Master's thesis, 2008 (A. Lux, C. Weidenbach)

Dilyana Dimova: On the Translation of Timed Automata into First-order Logic, Master's thesis, 2009 (A. Fietzke, C. Weidenbach)

32.12 Dissertations, Habilitations, Offers, Awards

32.12.1 Dissertations

– Thomas Hillenbrand: *Superposition and Decision Procedures – Back and Forth*, April 24, 2009.

32.12.2 Offers for Faculty Positions

– Viorica Sofronie-Stokkermans: Faculty position at University of Manchester (declined).

32.12.3 Awards

– Thomas Hillenbrand: Winner of the UEQ division of the CADE ATP system competition 2007.
– Thomas Hillenbrand: Winner of the UEQ division of the CADE ATP system competition 2008.

32.13 Grants and Cooperations

AVACS – Automatic Verification and Analysis of Complex Systems

AVACS is a transregional collaborative research center (SFB Transregio) linking the sites Oldenburg, Freiburg and Saarbrücken and funded by the German Research Foundation (DFG). The center addresses the rigorous mathematical analysis of models of complex safety critical computerized systems, such as aircrafts, trains, cars, or other artifacts, whose failure can endanger human life.

Its goal is to raise the state of the art in automatic verification and analysis techniques from its current level, where it is applicable only to isolated facets (concurrency, time, continuous control, stability, dependability, mobility, data structures, hardware constraints, modularity, levels of refinement), to a level allowing a comprehensive and holistic verification of such systems. This involves investigating the interrelationships of a whole spectrum of models, ranging from classical non-deterministic transition systems to probabilistic, real-time, and hybrid system models.

– Starting date: January 2004.
– Duration: 12 years.
– Funding: DFG Transregional Collaborative Research Center.
– Max Planck Institute principal investigators: Viorica Sofronie-Stokkermans, Uwe Waldmann, Christoph Weidenbach.
– Partners: Albert-Ludwigs-Universität Freiburg, Carl von Ossietzky Universität Oldenburg, Universität des Saarlandes.

Feature Modelling

Feature models are a standard way to represent intelligent bills of materials. We have designed a variant of SPASS that is able to decide satisfiability and consequence properties for the translation of feature models into propositional and first-order logic. The tool is integrated into the feature modelling tool VEIA of the PROSTEP AG and currently evaluated for industrial usage.

– Starting date: February 2008.
– Duration: 6 months.

– Funding: PROSTEP IMP GmbH

– Christoph Weidenbach, Patrick Wischnewski, Christian Dressler.

32.14 Publications

Books

[1] F. Baader, S. Ghilardi, M. Hermann, U. Sattler, and V. Sofronie-Stokkermans, eds. *Complexity, Expressibility, and Decidability in Automated Reasoning (CEDAR'08)*, -, 2008. Informal proceedings.

[2] S. Ghilardi, U. Sattler, V. Sofronie-Stokkermans, and A. Tiwari, eds. *Automated Deduction: Decidability, Complexity, Tractability (ADDCT'07)*, -, 2007. Informal proceedings.

Journal articles and book chapters

[1] J. Freiheit and F. Zangl. Model-based user-interface management for public services. *Electronic Journal of e-Government*, 5(1):53–62, 2007.

[2] S. Jacobs and V. Sofronie-Stokkermans. Applications of hierarchical reasoning in the verification of complex systems. *Electronic Notes in Theoretical Computer Science*, 174(8):39–54, 2007.

[3] S. Jacobs and U. Waldmann. Comparing instance generation methods for automated reasoning. *Journal of Automated Reasoning*, 38(1/3):57–78, 2007.

[4] V. Sofronie-Stokkermans. Algebraic and logical methods in computer science: some aspects. In A. Iorgulescu, S. Marcus, S. Rudeanu, and D. Vaida, eds., *Grigore C. Moisil and his followers*, pp. 488–493. Editura Academiei Romane, Bucharest, Romania, 2007.

[5] V. Sofronie-Stokkermans. Automated theorem proving by resolution in non-classical logics. *Annals of Mathematics and Artificial Intelligence*, 49(1-4):221–252, 2007.

[6] V. Sofronie-Stokkermans. On unification for bounded distributive lattices. *ACM Transactions on Computational Logic*, 8(2):12.1–12.28, 2007.

[7] V. Sofronie-Stokkermans. Interpolation in local theory extensions. *Logical Methods in Computer Science*, 4(4):31 pages, 2008. Special issue of LMCS dedicated to IJCAR 2006.

[8] V. Sofronie-Stokkermans. Sheaves and geometric logic and applications to modular verification of complex systems. *Electronic Notes in Theoretical Computer Science*, 230:161–187, 2009.

[9] V. Sofronie-Stokkermans and C. Ihlemann. Automated reasoning in some local extensions of ordered structures. *Journal of Multiple-Valued Logic and Soft Computing*, 13(4-6):397–414, 2007.

[10] D.-K. Tran, C. Ringeissen, S. Ranise, and H. Kirchner. Combinations of convex theories: Modularity, deduction completeness and explanation. *Journal of Symbolic Computation*, 2009. Accepted for publication.

[11] J. Ziemann, T. Matheis, and J. Freiheit. Modelling of cross-organizational business processes - current methods and standards. *Enterprise Modelling and Information Systems Architectures*, 2(2):23–31, 2007.

Conference articles

[1] P. Baumgartner and U. Waldmann. Superposition and model evolution combined. In R. Schmidt, ed., *22nd International Conference on Automated Deduction, CADE-22*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer.

[2] W. Damm, S. Disch, H. Hungar, S. Jacobs, J. Pang, F. Pigorsch, C. Scholl, U. Waldmann, and B. Wirtz. Exact state set representations in the verification of linear hybrid systems with large discrete state space. In K. S. Namjoshi, T. Yoneda, T. Higashino, and Y. Okamura, eds., *Automated Technology for Verification and Analysis, 5th International Symposium, ATVA 2007*, Tokyo, Japan, 2007, *LNCS 4762*, pp. 425–440. Springer.

[3] J. Faber, S. Jacobs, and V. Sofronie-Stokkermans. Verifying CSP-OZ-DC specifications with complex data types and timing parameters. In J. Davies and J. Gibbons, eds., *Proceedings of IFM 2007: Integrated Formal Methods*, Oxford, UK, 2007, *LNCS 4591*, pp. 233–252. Springer.

[4] A. L. Fietzke and C. Weidenbach. Labelled splitting. In *IJCAR*, Sydney, Australia, 2008, *LNCS 5195*, pp. 459–474. Springer.

[5] M. Horbach and C. Weidenbach. Superposition for fixed domains. In *CSL*, Bertinoro, Italy, 2008, *LNCS 5213*, pp. 293–307. Springer.

[6] M. Horbach and C. Weidenbach. Decidability results for saturation-based model building. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Computer Science. Springer. Accepted.

[7] C. Ihlemann, S. Jacobs, and V. Sofronie-Stokkermans. On local reasoning in verification. In C. R. Ramakrishnan and J. Rehof, eds., *Proceedings of TACAS 2008*, Budapest, Hungary, 2008, *LNCS 4963*, pp. 265–281. Springer.

[8] C. Ihlemann and V. Sofronie-Stokkermans. System description: H-PILoT version 1.5. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer.

[9] S. Jacobs. Incremental instance generation in local reasoning. In F. Baader, S. Ghilardi, M. Hermann, U. Sattler, and V. Sofronie-Stokkermans, eds., *Workshop: Complexity, Expressibility, and Decidability in Automated Reasoning - CEDAR'08*, Sydney, Australia, 2008, pp. 47–62.

[10] S. Jacobs. Incremental instance generation in local reasoning. In A. Bouajjani and O. Maler, eds., *Computer Aided Verification - 21st International Conference, CAV 2009*, Grenoble, France, 2009. Springer. To appear.

[11] T. Lev-Ami, C. Weidenbach, T. Reps, and M. Sagiv. Labelled clauses. In *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007, *LNCS 4603*, pp. 311–327. Springer.

[12] M. Ludwig and U. Waldmann. An extension of the Knuth-Bendix ordering with LPO-like properties. In N. Dershowitz and A. Voronkov, eds., *Logic for Programming, Artificial Intelligence, and Reasoning, 14th International Conference, LPAR 2007*, Yerevan, Armenia, 2007, *LNAI 4790*, pp. 348–362. Springer.

[13] C. Lynch and D.-K. Tran. Smels: Satisfiability modulo equality with lazy superposition. In S. S. Cha, J.-Y. Choi, K. Moonzoo, I. Lee, and M. Viswanathan, eds., *Automated Technology for Verification and Analysis 6th International Symposium, ATVA 2008*, Seoul, Korea, 2008, *LNCS 5311*, pp. 186–200. Springer.

[14] A. Rybalchenko and V. Sofronie-Stokkermans. Constraint solving for interpolation. In B. Cook and A. Podelski, eds., *Verification, Model Checking and Abstract Interpretation, 8th International Conference, VMCAI 2007*, Nice, France, 2007, *LNCS 4349*, pp. 346–362. Springer.

[15] V. Sofronie-Stokkermans. Hierarchical and modular reasoning in complex theories: The case of local theory extensions. In B. Konev and F. Wolter, eds., *Frontiers of Combining Systems. 6th International Symposium FroCos 2007, Proceedings*, Liverpool, UK, 2007, *LNCS 4720*, pp. 47–71. Springer. Invited paper.

[16] V. Sofronie-Stokkermans. Hierarchical and modular reasoning in complex theories: The case of local theory extensions. In S. Ranise, ed., *Proceedings of the Sixth International Workshop on First-Order Theorem Proving (FTP 2007)*, Liverpool, UK, 2007, *Technical Report, Department of Computer Science, University of Liverpool.*, vol. ULCS-07-018, p. 1. University of Liverpool. the full paper appeared in the proceedings of FroCos 2007.

[17] V. Sofronie-Stokkermans. On unification in certain finitely generated varieties of algebras. In E. Contejean, ed., *Proceedings of the 21th International Workshop on Unification (UNIF 2007)*, Paris, France, 2007, pp. 1–5.

[18] V. Sofronie-Stokkermans. Efficient hierarchical reasoning about functions over numerical domains. In K. Berns and T. Breuel, eds., *KI 2008: Advances in Artificial Intelligence*, Kaiserslautern, Germany, 2008, *LNAI 5243*, pp. 135–143. Springer.

[19] V. Sofronie-Stokkermans. Locality and subsumption testing in $\mathcal{EL}$ and some of its extensions. In C. Areces and R. Goldblatt, eds., *Advances in Modal Logic, Vol.7 (Proceedings of AIML 2008)*, Nancy, France, 2008, pp. 315–339. College Publications.

[20] V. Sofronie-Stokkermans. Locality and subsumption testing in $\mathcal{EL}$ and some of its extensions. In F. Baader, C. Lutz, and B. Motik, eds., *Proceedings of the 21st International Workshop on Description Logics (DL-2008)*, Dresden, Germany, 2008, p. 11pages. CEUR Workshop Proceedings.

[21] V. Sofronie-Stokkermans. Locality results for certain extensions of theories with bridging functions. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer. To appear.

[22] V. Sofronie-Stokkermans and C. Ihlemann. Automated reasoning in some local extensions of ordered structures. In *Proceedings of ISMVL 2007*, Oslo, Norway, 2007, p. Article1. IEEE.

[23] V. Sofronie-Stokkermans, C. Ihlemann, and S. Jacobs. Local theory extensions, hierarchical reasoning and applications to verification. In F. Baader, B. Cook, J. Giesl, and R. Nieuwenhuis, eds., *Deduction and Decision Procedures*, Dagstuhl, Germany, 2007, *Dagstuhl Seminar Proceedings*, vol. 07401, pp. 1–22. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

[24] C. Weidenbach, D. Dimova, A. Fietzke, M. Suda, and P. Wischnewski. Spass version 3.5. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Computer Science. Springer. Accepted.

[25] C. Weidenbach, R. Schmidt, T. Hillenbrand, R. Rusev, and D. Topic. System description: Spass version 3.0. In F. Pfenning, ed., *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007, *LNAI 4603*, pp. 514–520. Springer.

[26] C. Weidenbach and P. Wischnewski. Contextual rewriting in spass. In *PAAR/ESHOL*, Sydney, Australien, 2008, *CEUR Workshop Proceedings*, vol. 373, pp. 115–124. CEUR-WS.org.

[27] J. Ziemann, T. Matheis, and J. Freiheit. Modelling of cross-organizational business processes. In *Enterprise Modelling and Information Systems Architectures 2007*, St. Goar, Germany, 2007, pp. 87–95. SIG MoBIS.

Tutorials

[1] V. Sofronie-Stokkermans. Reasoning in complex theories and applications, 2008.

Theses

[1] D. Dimova. Propositional abduction. Bachelor thesis, Universität des Saarlandes, 2007.

[2] C. Dressler. First-order proof documentation. Bachelor thesis, Universität des Saarlandes, 2007.

[3] A. L. Fietzke. Labelled splitting. Masters thesis, Universität des Saarlandes, 2007.

[4] T. Hillenbrand. *Superposition and Decision Procedures – Back and Forth.* Phd thesis, Universität des Saarlandes, 2008.

[5] M. Lamotte. Analysis of Authorizations in SAP R/3. Masters thesis, Fachhochschule Trier, 2008.

[6] R. Rusev. Bitvector reasoning with spass. Masters thesis, Universität des Saarlandes, 2008.

[7] P. Wischnewski. Contextual rewriting in spass. Masters thesis, Universität des Saarlandes, 2007.

[8] S. Zimmer. Intelligent combination of a first order theorem prover and smt procedures. Diploma thesis, Universität des Saarlandes, 2007.

Technical reports

[1] A. Fietzke and C. Weidenbach. Labelled splitting. Research Report MPI-I-2008-RG1-001, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, 2008.

[2] T. Hillenbrand and C. Weidenbach. Superposition for finite domains. Research Report MPI-I-2007-RG1-002, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, 2007.

[3] S. Hirth, C. Karl, and C. Weidenbach. Automatic analysis of LAN infrastructures. Research Report MPI-I-2007-RG1-001, Max Planck Institute for Informatics, Saarbruecken, Germany, 2007.

[4] V. Sofronie-Stokkermans. Efficient hierarchical reasoning about functions over numerical domains. Reports of SFB/TR 14 AVACS ATR 45, SFB/TR 14 AVACS, SFB/TR 14 AVACS, 2008. extended version of an article with the same name published in the proceedings of KI 2008.

[5] V. Sofronie-Stokkermans. Sheaves and geometric logic and applications to modular verification of complex systems. Reports of SFB/TR 14 AVACS ATR 46, SFB/TR 14 AVACS, SFB/TR 14 AVACS, 2008.

[6] D.-K. Tran, C. Ringeissen, S. Ranise, and H. Kirchner. Combinations of convex theories: Modularity, deduction completeness and explanation. Research Report MPI-I-2008-RG1-002, Max-Planck-Institut for Informatics, MPI-INF, Campus E 1 4, 66123 Saarbrücken, 2009.

[7] C. Weidenbach and P. Wischnewski. Contextual rewriting. Research Report MPI-I-2009-RG1-002, Max-Planck-Institut für Informatik, MPI für Informatik, Campus E 1 4, 66123 Saarbrücken, 2009.

33 The Machine Learning Group (RG2)

33.1 Personnel

Head of the Group
Prof. Dr. Tobias Scheffer (January 2007–September 2008)

Secretary
Ellen Fries

PhD Students
Steffen Bickel (January 2007–September 2008)
Ulf Brefeld (January 2007–September 2007)
Michael Brückner (January 2007–September 2008)
Uwe Dick (January 2007–September 2008)
Laura Dietz (January 2007–)
Jochen Fischer (April 2008–September 2008, external PhD student)
Peter Haider (January 2007–September 2008)

Research Assistants
Thoralf Klein (January 2007–June 2008)
Barbara Pogorzelska (June 2007–November 2008)
Christoph Sawade (Mai 2007–October 2008)
Peter Siemen (January–December 2007)
Arvid Terzibaschian (July 2007–December 2008)

33.2 Group Organization

The group met weekly for the "Machine Learning Journal Club" on Mondays, 3 pm. During this meeting, either external researchers or group members gave talks on their current work and discussed relevant papers.

33.3 Research Projects

33.3.1 Transfer Learning

Investigators: Steffen Bickel and Michael Brückner

Most machine learning algorithms are constructed under the assumption that the training data is governed by the exact same distribution which the model will later be exposed

to. In practice, control over the training data is often less perfect. Training data may be obtained under laboratory conditions that cannot be expected after deployment of a system; spam filters may be used by individuals whose distribution of inbound emails diverges from the distribution reflected in public training corpora (*e.g.,* the TREC spam corpus); image processing systems may be deployed to foreign countries where vegetation and lighting conditions result in a distinct distribution of input patterns.

Learning under covariate shift refers to learning problems in which the marginal input distributions differ between training and test sets. In multi-task learning, multiple related learning problems have to be solved. The joint distributions of input and output differs across the tasks, labeled samples may be available for each task.

Discriminative Learning under Covariate Shift

In discriminative learning tasks such as classification, the classifier's goal is to produce the correct output given the input. This is best performed by learners that directly maximize a measure of the quality of the produced output.

Work in this line of research has contributed a discriminative model for learning under arbitrarily different training and test input distributions [2]. The model directly characterizes the divergence between training and test distribution, without the intermediate – intrinsically model-based – step of estimating training and test distribution. The search for all model parameters is formulated as an integrated optimization problem. This complements the predominant heuristic of first estimating the bias of the training sample, and then learning the classifier on a weighted version of the training sample. The integrated optimization problem can be maximized with a conjugate gradient procedure, leading to a kernel logistic regression classifier for covariate shift.

Work of the group has also contributed to a different model that is based on minimization of the KL divergence between test and weighted training distribution [4].

Multi-Task Learning

In multi-task learning one seeks to solve many classification problems in parallel. Some of the classification problems will likely relate to one another, but one cannot assume that the tasks share a joint conditional distribution of the class label given the input variables. The challenge of multi-task learning is to come to a good generalization across tasks: each task should benefit from the wealth of data available for the entirety of tasks, but the optimization criterion needs to remain tied to the individual task at hand.

A motivating application is the problem of predicting the therapeutic success of a given combination of drugs for a given strain of the Human Immunodeficiency Virus-1 (HIV-1).

We contribute a multi-task learning model that can handle arbitrarily different data distributions for different tasks without making assumptions about the data generation process or the relation between tasks. [1] We show that by appropriately weighting each instance in the pool of all examples, one can match the distribution that governs the pool of examples of all tasks to each of the single task distributions. We show Multi-Task Learning for HIV Therapy Screening how appropriate weights can be obtained by discriminating the labeled sample for a given task against the pooled sample.

Multi-Task Learning under Covariate Shift

This work addresses the problem of learning classifiers for several related tasks that may differ in their joint distribution of input and output variables. For each task, small — possibly even empty -– labeled samples and large unlabeled samples are available. While the unlabeled samples reflect the target distribution, the labeled samples may be biased. This setting is motivated by the problem of predicting sociodemographic features for users of web portals, based on the content which they have accessed. Here, questionnaires offered to a portion of each portal's users produce biased samples. We derive a transfer learning procedure that produces resampling weights which match the pool of all examples to the target distribution of any given task. Transfer learning enables us to make predictions even for new portals with few or no training data and improves the overall prediction accuracy [3].

References

- [1] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. Multi-task learning for hiv therapy screening. In W. W. Cohen, A. McCallum, and S. T. Roweis, eds., *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, Helsinki, Finland, 2008, ACM International Conference Proceedings Series 307, pp. 56–63. ACM Press.
- [2] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distribution. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, Corvallis, Oregon, USA, 2007, pp. 81–88. ACM Press.
- [3] S. Bickel, C. Sawade, and T. Scheffer. Transfer learning by distribution matching for targeted advertising. In *Advances in Neural Information Processing Systems 21, Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2009. MIT.
- [4] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. In *Proceedings of the SIAM International Conference on Data Mining*, Atlanta, USA, 2008, pp. 443–454. SIAM-Society for Industrial and Applied Mathematics.

33.3.2 Structural Learning

Investigators: Ulf Brefeld, Peter Haider, and Thoralf Klein

Learning mappings between arbitrary structured and interdependent input and output spaces covers many natural learning tasks such as producing sequential or tree-structured outputs, and it challenges the standard model of learning a mapping from independently drawn instances to a small set of labels. Potential applications include named entity recognition and information extraction (sequential output), natural language parsing (tree-structured output), classification with a class taxonomy – here, the output is a node in a tree –, and collective classification where the output is a set of interdependent class variables.

To capture the involved multiple-way dependencies, it is helpful to represent input and output pairs in a joint feature representation $\Phi(\mathbf{x}, \mathbf{y})$. The learning task aims at finding a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\bar{\mathbf{y}} \in \mathcal{Y}} f(\mathbf{x}, \bar{\mathbf{y}})$$

is the desired output for any input $\mathbf{x}$. Max-margin Markov models [5], support vector machines for structured output spaces [6] and other discriminative learners exploit this principle.

Transductive Learning for Structured Variables

In structured domains, labeled examples are frequently scarce while unclassified inputs are readily available. The question arises how unlabeled data can be effectively utilized by discriminative learners.

We have investigated transductive kernel machines for structured output variables [7]. In order to scale transductive learning to structured outputs, the corresponding non-convex, combinatoric, constrained optimization problems can be transformed into continuous, unconstrained optimization problems [1]. The discrete optimization parameters are eliminated and the resulting differentiable problems can be optimized more efficiently.

Approximate Inference for Structured Learning

In structural learning, a multi-variate prediction model predicts many dependent output variables simultaneously. The dependencies between the outputs give rise to complex inference problems that become intractable for large graphs. We have studied exact inference based on the junction tree algorithm and the approximate loopy belief propagation technique as inference procedure in learning problems for structural prediction [4].

Supervised Clustering of Streaming Data

This project addresses the challenge of clustering items in a data stream in a prescribed manner. It is inspired by the problem of detecting email batches in a mail transfer agent. Senders of spam, phishing, and virus emails do not send multiple identical copies of a message because later copies could be blocked based on blacklisting once the message is known to be malicious. Instead, they generate the individual messages by instantiating probabilistic grammars. In order to recognize batches, has to identify which messages have been jointly generated by the same grammar.

This problem naturally fits into the framework of structured learning [2], because the input is a set of messages and the output is an adjacency matrix on these elements. Training data consists of batches of messages together with the correct adjacency matrices. The learning problem amounts to finding parameters of a similarity function defined between pairs of emails such that a clustering based on this similarity identifies the training batches correctly.

Based on the nature of the data stream, the decoding – *i.e.,* clustering – procedure has to require at most linear time and the similarity function has to be found such that the linear decoder finds the correct adjacency matrices. Empirical results show that using collective features of entire email batches can improve the identification of spam, phishing and virus emails substantially [3]. This work has been distinguished with the best Student Paper Award of the International Conference on Machine Learning.

References

[1] O. Chapelle. Training a support vector machine in the primal. Technical Report 147, Max Planck Institute Tübingen, 2006.

[2] T. Finley and T. Joachims. Supervised clustering with support vector machines. In *Proceedings of the Twenty-Second International Conference on Machine Learning*, 2005.

- [3] P. Haider, U. Brefeld, and T. Scheffer. Supervised clustering of streaming data for email batch detection. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning (ICML 2007*, Corvallis, Oregon, USA, 2007, pp. 345–352. ACM Press.

- [4] T. Klein, U. Brefeld, and T. Scheffer. Exact and approximate inference for annotating graphs with structural svms. In W. Daelemans, B. Goethals, and K. Morik, eds., *Maschine Learning and Knowledge Discovery in Databases, Processings Part I*, Antwerp, Belgium, 2008, *LNCS 5211*, pp. 611–623. Springer.

[5] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems*, 2004.

[6] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

- [7] A. Zien, U. Brefeld, and T. Scheffer. Transductive support vector machines for structured variables. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, Corvallis, Oregon, USA, 2007, pp. 1183–1190. ACM Press.

33.3.3 Adversarial Learning and Information Security

Investigators: Michael Brückner, Peter Haider, and Uwe Dick

Most research on machine learning – and in fact all research in statistics – relies on the assumption that *nature* does not actively resist attempts to model its behavior. Many security-related applications require learning algorithms to model the behavior of *adversaries* who reverse-engineer any filtering or detection mechanism employed. Attackers specifically engineer their software tools such as to maximize the odds of successfully deceiving security mechanisms, thus creating a moving target for the learning algorithm employed [5, 7].

In an adversarial prediction game, a learner produces a classifier while an adversary may alter the input distribution. We have studied adversarial prediction games in which the loss functions of learner and adversary are negatives of each other, and the more general case of loss functions that are not necessarily directly antagonistic. For both cases, we have been able to identify conditions under which the prediction game has a unique Nash equilibrium, and derive algorithms that will find the equilibrial prediction model [3, 4].

Our investigations in this area are conducted in cooperation with the European web hosting company STRATO AG. Our work has led to tools that protect STRATO's servers and customers against spam, phishing, and virus emails [2], and protect STRATO against abuse of their infrastructure. We have conducted case studies on Nash-equilibrial prediction models in cooperation with STRATO.

Security application have also motivated parts of our work on personalized information filtering [1], and detection of batches [6].

References

- [1] S. Bickel and T. Scheffer. Dirichlet-enhanced spam filtering based on biased samples. In B. Schölkopf, J. Platt, and T. Hoffman, eds., *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada, 2007, vol. 19. MIT Press.

[2] M. Brückner, P. Haider, , and T. Scheffer. Highly scalable discriminative spam filtering. In *Proceedings of the Text Retrieval Conference (TREC)*, 2006.

- [3] M. Brückner and T. Scheffer. Optimal spamming: Solving a family of adversarial classification games. In *Proceedings of NIPS 2007 Workshop on Machine Learning in Adversarial Environments for Computer Security*, La Jolla, CA, 2007, pp. 1–2. NIPS Foundation.

[4] M. Brückner and T. Scheffer. Nash equilibria of adversarial prediction games. Unpublished manuscript, 2009.

[5] P. Domingos, N. Dalvi, Mausam, S. Sanghai, and D. Verma. Adversarial classification. In *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining*, 2004.

- [6] P. Haider, U. Brefeld, and T. Scheffer. Supervised clustering of streaming data for email batch detection. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning (ICML 2007*, Corvallis, Oregon, USA, 2007, pp. 345–352. ACM Press.

[7] D. Lowd and C. Meek. Adversarial learning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

33.3.4 Modeling Citation Influences

Investigators: Laura Dietz and Steffen Bickel

The goal in this line of research is to understand how machine learning technologies can help to construct systems that satisfy a user's information need better.

In order to obtain an understanding of a research field, a scientist has to identify papers that describe key contributions, understand how the various contributions relate to one another, and how the topic evolved over the past. In principle, publication repositories contain all the information required to accomplish this. But visualization of, and search in, the space of scientific results are still unsolved problems.

Google Scholar, CiteSeer and other tools that allow to navigate in the publication graph have made this task much easier. But still these tools operate on the syntactic space of *publications* rather than the semantic space of *results* and *knowledge*. Making it possible to search and to visualize the latent structure, topics, key contributions and the flow of results between papers remains a vision that motivates research in this area.

The citation influence model (Figure 33.1) explains the generation of documents; the model incorporates the aspects of topical innovation and topical inheritance via citations. The model has been implemented in a citation influence browser that visualizes the impact that research papers have had on one another. It is based on a probabilistic, generative model that characterizes topics as parameters of the distributions of their textual manifestations. Model parameters are estimated from publication repositories using Markov Chain Monte Carlo techniques. The model produces graphical visualizations of research topics in papers and the strength of influences between papers [1, 2, 3].

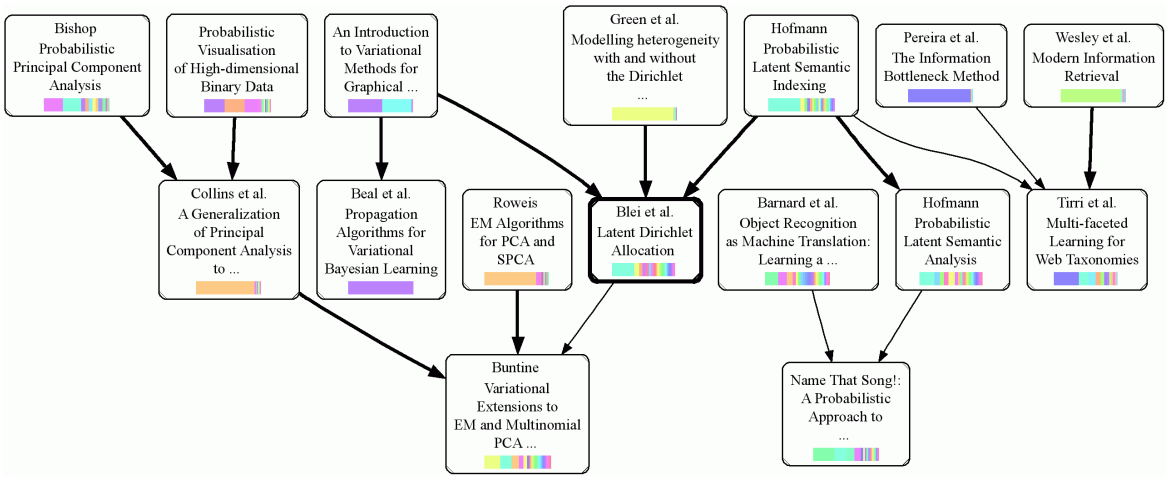


Figure 33.1: Influence strengths of links in the citation vicinity of the publication "Latent Dirichlet Allocation". Strength of arcs represents the strength of influence; colors indicate topics.

References

[1] L. Dietz. Probabilistic topic models to support a scientific community. In *Twenty-First National Conference on Artificial Intelligence*, 2006.

- [2] L. Dietz and S. Bickel. Modeling evolution of ideas in the web of science. In *Proceedings of NIPS'07 workshop on statistical models of networks*, Whistler, Canada, 2007, pp. 1–2.

- [3] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, Corvallis, Oregon, USA, 2007, pp. 233–240. ACM Press.

33.3.5 Knowledge Discovery from Streams

Investigator: Tobias Scheffer

Many data streams have too high a volume for any knowledge discovery algorithm to process them in their entirety. In cooperation with Szymon Jaroszewicz, Polish Institute of Telecommunication and Lenka Ivantysynova of Humboldt-Universität zu Berlin we investigate algorithms that learn from infinite streams efficiently and can yet be guaranteed to come ε-close to the solution that would be attained if all (infinitely many) data were to be processed. We have derived such algorithms for several cases [3, 3]. In some interesting application scenarios, background knowledge is available in a Bayesian network. Integrating this background knowledge into the discovery process leads to algorithms that find the most surprising, unexpected patterns, rather than patterns that are already well-known. The Apriori-BNS algorithm is able to handle arbitrarily large databases and Bayesian networks that are too large for exact inference to be feasible [2]. Similar near-optimality can be obtained for the problem of matching schemas of data streams [1].

References

- [1] S. Jaroszewicz, L. Ivantysynova, and T. Scheffer. Schema matching on streams with accuracy guarantees. *Intelligent Data Analysis*, 12:253–270, 2008.
- [2] S. Jaroszewicz, T. Scheffer, and D. A. Simovici. Scalable pattern mining with bayesian networks as background knowledge. *Data Mining and Knowledge Discovery*, 2008.

[3] T. Scheffer and S. Wrobel. Finding the most interesting patterns in a database quickly by using sequential sampling. *Journal of Machine Learning Research*, 3:833–862, 2002.

33.4 Academic Activities

33.4.1 Journal Positions

Tobias Scheffer serves

- as an Action Editor of the *Data Mining and Knowledge Discovery Journal;*
- on the Editorial Board of the *Machine Learning Journal;*
- on the Editorial Board of the *Journal of Artificial Intelligence Research.*

33.4.2 Conference and Workshop Positions

Conference Organization

Tobias Scheffer serves as

- General Chair of the *Conference on Email and Anti-Spam 2009.*
- Program Co-Chair of the *Conference on Email and Anti-Spam 2008.*

Membership in Steering Committee

Tobias Scheffer serves on the Steering Committee of the

- *European Conference on Machine Learning.*
- *European Conference on Principles and Practice of Knowledge Discovery in Databases.*
- *International Conference on Discovery Science.*

Membership in Program Committees

Members of the group serve on the following Program Committees.

Steffen Bickel

- *NIPS 2008 Workshop on Cost Sensitive Learning.*
- *European Conference on Machine Learning* 2007, 2008.
- *European Conference on Principles and Practice of Knowledge Discovery in Databases* 2007, 2008.
- *International Conference on Machine Learning.* 2007.

– *Pacific-Asian Conference on Knowledge Discovery in Databases*. China. 2007.

Laura Dietz

– *European Conference on Machine Learning* 2008, 2009.
– *ESCW Workshop on Bridging the Gap between Semantic Web and Web 2.0*. 2007.

Tobias Scheffer

– *International Conference on Machine Learning* 2007, 2008, 2009.
– *ACM Conference on Knowledge Discovery and Data Mining* 2008, 2009.
– *International Joint Conference on Artificial Intelligence* (SPC) 2009.
– *European Conference on Machine Learning* (Area Chair), 2007, 2008.
– *European Conference on Principles and Practice of Knowledge Discovery in Databases* (Area Chair). 2007, 2008.

33.4.3 Reviewing for Funding Organizations

Tobias Scheffer served as reviewer for the following funding agencies.

– *National Science Foundation.*
– *German Science Foundation DFG.*
– *Czech Science Foundation.*
– *Belgian Science Foundation.*

33.4.4 Invited Talks and Tutorials

Tobias Scheffer:

– *Learning from Multiple Sources by Matching their Distributions*. Invited talk at the NIPS 2008 Workshop on Learning from Multiple Data Sources.
– *Empfehlungsalgorithmen für Web 2.0-Anwendungen*, Data Mining Anwendertage, 2008.
– *Machine Learning for Recommender Systems*. Bellairs Workshop on Program Analysis for Recommendation Systems, 2008.
– *Transfer Learning*. Spring Workshop on Mining and Learning, 2008.
– *Learning under Differing Training and Test Distributions*, Johns Hopkins University, 2007.
– *Den gläsernen Menschen bewirtschaften: Online-Werbung, Spam, Online-Betrug*, Vortragsreihe *Der gläserne Mensch* der Stadt Saarbrücken, 2007.
– *Email Security: Detecting Spam, Phishing, Intrusions*, NIPS Workshop on Adversarial Learning, 2007.
– *Phishing, Pharming, Phraud*. Invited Talk. Annual Conference of the German Society for Classification (GfKL). Freiburg. 2007.
– *Challenges of Structured Prediction*. Keynote at FLUFFY 2007.

- *Challenges of Natural Language Processing.* Keynote at the Annual Meeting of the IRTG in Language Technology and Cognitive Systems. 2007.
- *Maschinelles Lernen für knifflige Sprachverarbeitungsprobleme.* Albert-Ludwigs-Universität Freiburg. 2007.

33.5 Dissertations, Habilitations, Offers, Awards

33.5.1 Completed Dissertation Projects

- Steffen Bickel: Learning for differing training and test distributions, 2008.
- Ulf Brefeld: Semi-supervised learning for structured variables, 2007.

33.5.2 Offers for Faculty Positions

- Tobias Scheffer: W3 position at the Albert-Ludwigs-Universität Freiburg, 2007.
- Tobias Scheffer: W3 position at Technische Universität Chemnitz, 2007.
- Tobias Scheffer: W3 position at Universität Potsdam, 2008.

33.5.3 Awards

- Peter Haider, Ulf Brefeld: Best Student Paper Award at the International Conference on Machine Learning, 2007.
- Tobias Scheffer: Google Research Award, 2008.

33.6 Grants and Cooperations

33.6.1 Differing Training and Test Distributions in Active Learning

Funding: Google Research Award

Duration: 2009.

Project Members: Christoph Sawade.

Active learning reduces the labeling effort incurred by applying machine learning algorithms. Active learning procedures direct the attention of a labeler towards examples whose label is believed to convey a maximum of information. Labeled samples in active learning are governed by a distribution that differs from the natural test distribution for multiple reasons. An initial labeled sample may be compiled from auxiliary data sources; the natural input distribution may change over time, or may be altered by an adversary. In addition – and specific to active learning – an active instance selection procedure creates a labeled sample that is biased by the selection criterion. Treating the artificially selected sample in active learning as if it was governed by the test distribution is not necessarily the best course of action. We will understand, develop, and evaluate systematic approaches to active learning that account for this discrepancy between labeled training and test distributions.

33.6.2 Mining Jazz Data to Assess Development Processes

Funding: IBM Jazz Faculty Grant to Andreas Zeller (Saarland University) and Tobias Scheffer

Duration: 2008-2009.

What is it that makes a good development process? We want to develop a plug-in that learns from collaboration and defect data as tracked by Jazz, relates features of the collaborative development process to the defect density of individual components, and thereby automatically predicts code quality. For instance, the plug-in might advise that package P should be reviewed more, because a new dependency on compiler internals has been added shortly before the release date by a developer who is new to the team.

33.6.3 Modelling and Optimization of Dialysis Treatment

Funding: Fresenius NephroCare e-Services GmbH

Duration: 2008-2010.

Project Members: Jochen Fischer.

We investigate model-building and the generation of actionable knowledge from records of dialysis treatments.

33.6.4 Personalized Ranking of Online Advertisements

Funding: nugg.ad AG

Duration: 2007-2008.

Project Members: Christoph Sawade, Arvid Terzibaschian.

In this project, we investigate efficient algorithms that predict which advertisement a user is most likely to click at, based on that user's past clicking behavior and all other information that is available.

33.6.5 Data and Text Mining in Quality and Service

Funding: Daimler AG

Duration: 08/2005-09/2008

Project Members: Peter Siemen

We study the problem of discovering trends and new developments in production and warranty databases as well as in workshop reports. We develop technologies that automatically identify such trends and discover their hidden causes. The goal of this project is the constructive analysis of data mining methods that lead to improved service processes by integrating and analyzing textual information and data from multiple, heterogeneous and distributed databases.

33.6.6 Email Security

Funding: Strato Rechenzentrum AG
Duration: since 07/2005
Project Members: Michael Brückner, Peter Haider, Uwe Dick.

We analyze a compound of problems that revolve around the security of email and web hosting services: containment of email spam, phishing, and virus outbreaks, and the detection of abuse of services and computing infrastructure. These applications involve adversarial prediction problems in which an active opponent modifies the input distribution in order to prevent the learner from building an accurate model. We study adversarial prediction problems, and develop technologies on a prototypical level that are continuously being integrated into the software that Strato employs to offer its services.

33.6.7 Text Mining: Knowledge Discovery in Text Databases and Efficient Document Processing

Funding: German Science Foundation DFG
Duration: June 2003 through December 2008
Project Members: Steffen Bickel, Ulf Brefeld, Thoralf Klein

The amount of documents available in archives and on the web is growing exponentially. This growth induces a demand for methods that automatically analyze large volumes of documents, discover and utilize valuable knowledge contained in them. A substantial part of our working processes consists of processing (i.e., reading, writing, manipulating) documents. Many tools support the administration of text documents, such as file systems, databases, or document management systems. Much greater efforts (and more expenses), however, are imposed by the actual document manipulation processes – such as writing documents. Any support of document manipulation processes requires substantial knowledge; it is therefore much more difficult to support document processing rather than document administration. The goal of the "Text Mining" project is to develop and study text mining algorithms that discover knowledge in large document archives, and utilize this knowledge to support future text manipulation processes.

33.6.8 Cooperations

Learning from Traces of Software Errors

The goal of this cooperation with the Software Engineering Group (Andreas Zeller) at the Saarland University, Department 5 of the Max Planck Institute for Informatics (Gerhard Weikum), and Matthias Hein of the Saarland University is to understand how learning algorithms can help to identify and locate software errors by analyzing various sources of data, such as databases of automatically generated bug reports and execution traces.

Machine Learning for Bioinformatics Problems

We have addressed plan to address selected machine learning problems for applications in bioinformatics in cooperation with the Department 3 (Thomas Lengauer). The prediction of

effectiveness of HIV drugs is hindered by the difference between training and test distribution, caused by the rapid development of the virus. New results on learning under covariate shift provide leverage to this problem. The prediction of protein interaction involves several challenges that have been addressed in the context of learning with structured output variables.

33.7 Publications

The following articles have been published between April 2007 and September 2008.

Journal Articles

[1] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, In print. In print.

[2] S. Jaroszewicz, L. Ivantysynova, and T. Scheffer. Schema matching on streams with accuracy guarantees. *Intelligent Data Analysis*, 12:253–270, 2008.

[3] S. Jaroszewicz, T. Scheffer, and D. A. Simovici. Scalable pattern mining with bayesian networks as background knowledge. *Data Mining and Knowledge Discovery*, 2008.

Book Chapter

[1] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift with a single optimization problem. In J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, eds., *Dataset Shift in Machine Learning*, ch. 9, pp. 161–177. MIT, Cambridge, USA, 2009.

Conference Articles

[1] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. Multi-task learning for hiv therapy screening. In W. W. Cohen, A. McCallum, and S. T. Roweis, eds., *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, Helsinki, Finland, 2008, ACM International Conference Proceedings Series 307, pp. 56–63. ACM Press.

[2] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distribution. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, Corvallis, Oregon, USA, 2007, pp. 81–88. ACM Press.

[3] S. Bickel, C. Sawade, and T. Scheffer. Transfer learning by distribution matching for targeted advertising. In *Advances in Neural Information Processing Systems 21, Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2009. MIT.

[4] M. Brückner, P. Haider, and T. Scheffer. Highly scalable discriminative spam filtering. In *Proceedings of the 15th Text Retrieval Conference (TREC)*, 2007.

[5] U. Dick, P. Haider, and T. Scheffer. Learning from incomplete data with infinite imputations. In W. W. Cohen, A. McCallum, and S. T. Roweis, eds., *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, Helsinki, Finland, 2008, ACM International Conference Proceedings Series 307, pp. 232–239. ACM Press.

[6] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, Corvallis, Oregon, USA, 2007, pp. 233–240. ACM Press.

[7] P. Haider, U. Brefeld, and T. Scheffer. Supervised clustering of streaming data for email batch detection. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning (ICML 2007*, Corvallis, Oregon, USA, 2007, pp. 345–352. ACM Press.

[8] D. S. Vogel, O. Asparouhov, and T. Scheffer. Scalable look-ahead linear regression trees. In P. Berkhin, R. Caruana, and X. Wu, eds., *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, USA, 2007, pp. 757–764. ACM.

[9] A. Zien, U. Brefeld, and T. Scheffer. Transductive support vector machines for structured variables. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, Corvallis, Oregon, USA, 2007, pp. 1183–1190. ACM Press.

Workshop Articles

[1] M. Brückner and T. Scheffer. Optimal spamming: Solving a family of adversarial classification games. In *Proceedings of NIPS 2007 Workshop on Machine Learning in Adversarial Environments for Computer Security*, La Jolla, CA, 2007, pp. 1–2. NIPS Foundation.

[2] L. Dietz. Probabilistic graph models for debugging software. In *Proceedings of NIPS 2008 Workshop on Analyzing Graphs: Theory and Applications*, Whistler, Canada, 2009, pp. 1–8.

[3] L. Dietz and S. Bickel. Modeling evolution of ideas in the web of science. In *Proceedings of NIPS'07 workshop on statistical models of networks*, Whistler, Canada, 2007, pp. 1–2.

Patent Disclosures

[1] S. Bickel, P. Haider, T. Scheffer, and R. Wienholtz. A computer implemented system and a method for detecting abuse of an electronic mail infrastructure in a computer network. European Patent Application EP07004097, 2007.

[2] P. Haider, A. Jansen, and T. Scheffer. A method of filtering electronic mail and an electronic mail system. European Patent Application EP07004098, 2007.

34 The Computational Genomics and Epidemiology Group (IRG1)

34.1 Personnel

Head of research group
Dr. Alice C. McHardy

Researchers
Dr. Ben Adams (September 2007–February 2008)

PhD Students
Sebastian Konietzny
Kaustubh Patil
Lars Steinbrück
Christina Tusche

System Administration
Dr. Joachim Büch

Secretary
Ruth Schneppen-Christmann

34.2 Visitors

In the time period from September 2007 to April 2009, the following researchers visited our group:

Dr. Gyan Bhanot	24.10.07	BioMaPS Institute and Biomedical Engineering, Rutgers University and Cancer Institute of New Jersey

34.3 Group Organization

The group meets in a weekly internal seminar, in which group members discuss the status of their projects, present their latest work or related topics and discuss and exchange ideas. Alice McHardy serves as a scientific mentor for the PhD students of the group. Furthermore, the group participates in the weekly seminar of Department 3, with group members presenting their work in this series approximately once a year.

34.4 Computational Methods for Metagenomics

34.4.1 Phylogenetic Classification of Variable-length DNA Sequences using Structural Support Vector Machines

Investigators: Kaustubh Patil and Alice McHardy

Metagenomics

Microbes are critical for the functioning of the ecosystem, agriculture, industry and also for the health of humans, other animals and plants. Thus understanding of microbes can be used to protect the environment, drug discovery and sustainable agriculture, among other applications [7]. The vast majority of microbes resist cultivation in a laboratory and conventional genome sequencing techniques cannot be used to study their diversity. Based on 16S ribosomal RNA (rRNA) surveys, it has been estimated that less than 1% of all Bacteria and Archaea can be cultivated [3]. Recent developments in the field of metagenomics (or environmental genomics) allow an unprecedented look into this hidden microbial world. Metagenome methods sequence randomly pooled DNA fragments from all microbes in an environmental sample. By sequencing naturally coexisting microbes, it is possible to explore microbial diversity and study the function of microbial communities in their natural habitat [3]. Metagenome sequencing generates a mix of short sequence fragments (between 30 and several hundred base pairs in length, depending on the sequencing technique) called reads. Overlapping reads can be assembled into longer sequence contigs. The contigs represent partial genome sequences of the sequenced organismal mixture. A further important step in metagenome sequence analysis is to deconvolute this fragment mixture into phylogenetic bins, representing the sampled organisms of a given community. The aim is to assign taxonomic information to these sequence fragments. In the following section we describe metagenomic binning in short, followed by our proposed method.

Metagenome Binning: State-of-the-art

The aim of a metagenome binning method is to assign taxonomic information to DNA sequence fragments [6]. As a gold standard 16S rRNA and other highly conserved genes can be used to phylotype DNA fragments. This method is very accurate but has very low coverage of genomic fragments and excludes many sequences from the analysis. Alternatively, sequence homology can be used on a less restricted set of genes to find similar sequences of known organisms, which in turn can be used to assign taxonomic identification. Generally, the success of this approach depends on the sequence coverage of a given part of the phylogenetic tree and in particular at larger evolutionary distances, accuracy may decrease due to processes such as gene loss, genesis and horizontal gene transfer. Instead of sequence conservation, it is also possible to use sequence composition, as oligonucleotide frequencies carry a phylogenetic signal. Various sequence composition-based supervised and unsupervised methods have been proposed for binning. Supervised methods use a reference set of genomic sequences along with their taxonomic information to build models. These models can then be used to infer the taxonomic origin for novel sequences. Unsupervised methods do not need reference

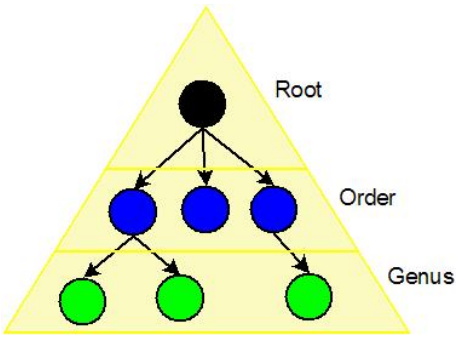


Figure 34.1: Example hierarchical output structure.

information[1]. PhyloPythia [5] is a supervised method based on hierarchically structured support vector machines (SVM) [9]. PhyloPythia exhibits state-of-the-art performance [4], but there is still some room for improvement, in particular for the assignment of short fragments and in the model building time. The following sections describe a novel approach, which we believe will show further improvements.

Structured Output Prediction

Recently, structured output prediction has gained widespread attention of the machine learning community and various approaches have been proposed [1]. Structured output prediction is different compared to the traditional classification problem in that the output (and also possibly the input) is more complex, containing structured dependencies between the different output values. Structural output methods exploit these dependencies in order to obtain more accurate models in less training time. One particular approach is based on the discriminative large-margin formulation [8] (hereafter referred to as structural SVM). We adopt this framework because of the strong theoretical foundations and empirical performance demonstrated by the support vector methods. The structural SVM is a generalization of multi-class SVM. Moreover, this model can be tuned with a loss function specific to the problem at hand. In the next section we briefly outline how this approach can be used for the phylogenetic binning task.

Structural SVM for Metagenomic Binning

For the binning task, the output is a hierarchical structure. Each node in the hierarchy represents a phylogenetic clade, higher level clades being more general. Figure 34.1 shows an example of a hierarchical output. We use the NCBI taxonomy [2] as a reference for the output structure. Publicly available genomes from GenBank [2] are used as reference data. The genomes were split into fragments of predefined lengths (3 KB, 5 KB, 10 KB, 15 KB and 50 KB) and oligonucleotide frequencies of length 4, 5 and 6 were extracted. These fragments along with their taxonomic assignment were then used to build structural SVM models.

[1] It should be noted that unsupervised methods do need some information in the assignment phase (e.g. 16S rRNA phylotyping).

Conclusions and Future Directions

Initial experiments with a structural SVM for metagenome binning show promising results. The results indicate that longer oligonucleotides have a higher discriminative power. Furthermore, classification accuracy correlates with fragment length. In the next phase of this work the method will be benchmarked on existing reference data sets [4]. Moreover, we are currently working on a feature extraction method which will allow to identify the oligonucleotide signatures of individual clades.

References

[1] G. Bakir, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan, eds. *Predicting structured data (Neural Information Processing)*. The MIT Press, 2007.

[2] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Research*, 37(suppl 1):D26–31, 2009.

[3] P. Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome Biology*, 3(2), 2002.

• [4] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6):495–500, 2007.

• [5] A. McHardy, H. G. Martin, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, 4(1):63–72, 2007.

• [6] A. McHardy and I. Rigoutsos. What's in the mix? Methods for the phylogenetic classification of metagenome sequence samples. *Current Opinion in Microbiology*, 10(5):499–503, 2007.

[7] C. on Metagenomics N.R.C. (U.S.). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. National Academies Press, 2007.

[8] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

[9] V. N. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, 1999.

34.4.2 Functional Module Inference and Process-level Genome Annotation

Investigators: Sebastian Konietzny, Laura Dietz (Department 5) and Alice McHardy

Motivation

Metagenomic sequencing has retrieved large numbers of novel genes from uncultivable organisms that are of biotechnological, medical or agricultural interest. For instance, 1,700 novel protein families were discovered in a comprehensive sequencing study of the marine planktonic microbiota from oceanic surface waters [1]. However, the functions and corresponding biological processes of these genes remain largely unknown, as they lack homology to experimentally characterized proteins of known function. We want to develop a method that processes large sets of (meta-) genome annotations and captures the inherent functional relationships between gene products acting in concert in a biological process by means of a probabilistic model. Such a model could then be used for a process-level annotation of genes,

a detailed characterization of the biological processes encoded in a (meta-) genome and to extend the current knowledge of biological processes.

References

[1] S. Yooseph, G. Sutton, D. B. Rusch, J. C. Venter, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biology*, 5(3):e16, 2007.

34.5 Computational Analysis of Influenza Evolution

34.5.1 Vaccine Strain Prediction

Investigators: Lars Steinbrück and Alice McHardy

Introduction

Influenza is one of the major health risks in modern life, responsible for up to 500,000 deaths annually [2]. Three distinct viral types (A, B and C) circulate in the human population. While types B and C evolve slowly and circulate at low levels, influenza A through rapid evolution continuously evades host immunity from previous infection or vaccination and regularly causes large epidemics. The virus is a single-stranded, negative-sense RNA virus of the family *Orthomyxoviridae*. The antigenic properties of a specific virus are defined by the surface glycoproteins, haemagglutinin (HA) and neuraminidase. These proteins are under ongoing selection for change and continuously accumulate mutations which result in reduced immune recognition in previously infected or vaccinated hosts – a process referred to as antigenic drift.

Motivation

While there are several drugs available for treatment of the disease (e.g. Tamiflu®), vaccination is the most effective method to prevent an influenza virus infection. The influenza vaccine is trivalent and includes one strain each for the currently circulating A/H1N1, A/H3N2 and B viruses. Due to antigenic drift, antigenically novel strains of influenza A appear and become predominant on a regular basis, which requires frequent adaptation of vaccine composition. To monitor for novel emerging strains, the World Health Organization (WHO) maintains a global surveillance program. A panel of experts meets twice a year to review the antigenic and genetic surveillance information about circulating strains and recommend the vaccine composition for each of the northern and southern hemispheres. Due to the time requirements of production and distribution, the recommendation has to be made almost one year before the season in which the vaccine is to be used. This 'predict and produce' approach mostly results in efficacious vaccines which substantially limit the morbidity and mortality of seasonal epidemics. However, problems arise when a novel emerging variant is detected too late to update the vaccine composition. Thus, for seasons preceding the establishment of a novel antigenic drift variant, increasing the prediction horizon through early and accurate detection is of the utmost importance [1]. In this project we aim to develop a method for recommending vaccine strains for influenza A/H3N2 (the more rapidly evolving subtype of endemic influenza

A) based on the genetic and epidemiological characteristics of sequenced isolates to improve the detection and control of emerging antigenic types.

References

[1] C. Russell, T. Jones, I. Barr, N. Cox, R. Garten, V. Gregory, I. Gust, A. Hampson, A. Hay, A. Hurt, et al. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, 26:31–34, 2008.

[2] K. Stohr. Influenza–WHO cares. *Lancet Infectious Diseases*, 2(9):517, 2002.

34.5.2 The role of year round and seasonal transmission paradigms in the evolution of influenza

Investigators: Ben Adams and Alice McHardy

Introduction

In temperate regions seasonal human influenza is known as a winter illness. In the northern hemisphere epidemics last from October to March, in the southern hemisphere from March to September. Infections are rarely observed in the respective summer months. Historically, influenza incidence in temperate regions has been estimated based on excess winter mortality relative to a summer baseline [3, 7]. The lack of such a baseline in non-seasonal environments makes tropical and subtropical incidence more difficult to quantify. However, recent studies have shown that influenza is present throughout the year in these parts of the world and, over the course of a whole year, the total number of infections is similar in equatorial and temperate regions [7].

Several theories have been proposed to explain how influenza survives the summer months in temperate regions. It has been suggested that it continues to circulate subclinically at low prevalence, remains in an inert state or becomes locally extinct and is re-seeded from the opposite hemisphere or the tropics [8]. Analysis of whole genome sequence data from New York state, Australia and New Zealand revealed evidence of external seeding. Multiple strains were shown to co-circulate each year, some appearing in opposite hemispheres for several consecutive seasons and others vanishing without trace after a single season. For all strains, however, there was little detectable evolution over the course of a single temperate transmission season [9]. Further genetic and antigenic analysis has subsequently identified East-Southeast Asia, where transmission is year round, as a source of global diversity. Novel strains appear in this region on average 6 - 9 months earlier than in Europe, North America and Oceania, and up to 12 - 18 months before appearing in South America [5].

Here we develop a suite of mathematical models to understand how seasonal and year round transmission regimes, along with the manner in which they are coupled, affect the epidemiology and evolution of influenza virus. We show that the boom-bust epidemiology associated with seasonal transmission limits the potential for an ongoing process of antigenic mutation within and across seasons, while the slower, steadier epidemiology associated with year round transmission facilitates the emergence of antigenic mutants locally and their subsequent global dissemination. Hence, fundamental epidemiological mechanisms indicate the importance of a region with established year round transmission for the continued persistence of influenza virus and may hold the key for the prediction and control of winter flu.

Mathematical Model

We consider global circulation, with two transmission paradigms, focusing on the 3 - 5 year period when a single antigenic cluster is dominant [4, 6], but making occasional excursions to investigate transitions in the prevalent cluster. For the seasonal paradigm, corresponding to temperate regions, the year is divided into transmission and non-transmission seasons, each lasting six months. During the transmission season the transmission rate parameter, which represents the probability that the virus is successfully passed on given contact between an infectious and a susceptible person, varies sinusoidally. The basic reproductive number, R_0, defined as the expected number of secondary infections resulting from a single infected individual in an immunologically naive population [2], is 1.0 at the beginning of the season, peaks after three months and declines symmetrically to the starting value after a total of six months. During the non-transmission season the transmission probability is zero. For the year round paradigm, the transmission rate parameter remains constant throughout the year. Infection incidence, however, will only be constant if the epidemiological dynamics have equilibrated to an endemic state. The observation of epochal changes in the dominant antigenic cluster of the influenza virus suggests that such stable incidence is unlikely and we expect mild epidemic surges associated with the punctuated appearance of escape mutants, possibly building gradually and spanning several years so less pronounced than the sharp seasonal epidemics of temperate regions.

These transmission paradigms are incorporated into the framework of a standard susceptible infected recovered model [1] with multiple co-circulating strains defined by 50-element binary 'genotypes'. Forty elements are antigenically neutral but the remaining ten determine the degree of cross-immunity, modeled in terms of the probability that a host previously infected with one strain is protected when challenged with another. Greater similarity between the antigenic elements of the bitstrings results in a higher probability of protection. We consider a suite of epidemiological models aimed at improving our understanding of how and where the influenza virus evolves. Initially we develop our understanding by considering each of the transmission paradigms separately, with one, two and multiple co-circulating strains. We then examine the impact of coupling regions with seasonal and year round transmission according to different paradigms. For example, by specifying that individuals in the northern region are exposed to a small proportion of infected individuals from the year round transmission patch.

References

[1] R. Anderson and R. M. May. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1991.

[2] F. Braur. *Basic ideas of mathematical epidemiology, Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An introduction*, vol. 125. Springer, 2002.

[3] M. I. Nelson and E. C. Holmes. The evolution of epidemic influenza. *Nature Reviews Genetics*, 8:196–205, 2007.

[4] J. B. Plotkin, J. Dushoff, and S. A. Levin. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proceedings of the National Academy of Science of the USA*, 99:6263–6268, 2002.

[5] C. A. Russell, T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, V. Gregory, I. D. Gust, A. W. Hampson, A. J. Hay, A. C. Hurt, J. C. de Jong, A. Kelso, A. I. Klimov, T. Kageyama, N. Komadina, A. S. Lapedes, Y. P. Lin, A. Mosterin, M. Obuchi, T. Odagiri, A. D. Osterhaus, G. F. Rimmelzwaan, M. W. Shaw, E. Skepner, K. Stohr, M. Tashiro, R. A. Fouchier, and D. J. Smith. The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320:340–346, 2008.

[6] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. Osterhaus, and R. A. Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305:371–376, 2004.

[7] C. Viboud, W. J. Alonso, and L. Simonsen. Influenza in tropical regions. *PLoS Medicine*, 3:e89, 2006.

[8] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka. Evolution and ecology of influenza A viruses. *Microbiological Reviews*, 56:152–179, 1992.

[9] Y. I. Wolf, C. Viboud, E. C. Holmes, E. V. Koonin, and D. J. Lipman. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology Direct*, 1:34, 2006.

34.6 Association Discovery in Genome-wide Association Studies

34.6.1 Mining for Subtle Disease-specific Patterns of SNP Markers and Associated Pathways

Investigators: Christina Tusche and Alice McHardy

In contrast to Mendelian (single-gene) diseases, complex diseases are caused by mutations in several genes, which are often interacting in unknown ways and furthermore influenced by various environmental factors. Examples for such conditions are cancer, Alzheimer's disease and common autoimmune diseases such as type-1 diabetes, rheumatoid arthritis and Crohn's disease. Genome-wide association studies (GWAS) identify associations between a large collection of genetic markers and a quantitative or qualitative trait with complex biological background, such as the status of a complex disease. The scientific field is highly multifaceted and on the way to overcome initial problems, such as the lack of statistical power due to small sample sizes and latent population substructures in the data. Disease-associated loci are commonly discovered based on single-marker or multi-marker statistical tests on single nucleotide polymorphisms (SNPs) in case-control data. A multitude of novel loci associated with complex diseases and related quantitative traits has thus been identified [1, 4].

Despite these impressive results, moving from the identification of association signals to the discovery of one or more *causal* genetic elements and, finally, to the underlying biological mechanism, remains one of the greatest challenges in GWAS [6]. The discovered loci individually account for a much smaller fraction of the population variation than estimates of the heredity of most common diseases [2], which leaves most of the genetic component for the disease unexplained. This might be due to the lack of power of the current studies for identification of variants with very small effect sizes as well as interactions between loci. Together with the development of techniques that can cope with the high dimensionality and intrinsic higher-order correlations, incorporating additional biological information in the search for association signals is of importance. In our project we plan to apply advanced

multi-locus models [5] and data mining techniques to mine for subtle patterns of SNPs in case-control data for complex diseases and to furthermore include information on functional interactions between genes in the discovery process. The benefit of multi-locus models and data mining techniques such as linear regression models, decision trees or random forests [8] is their flexibility. They allow the assessment of the joint (non-linear) impact of sets of genetic markers and the incorporation of further risk factors, such as environmental influences [3]. We believe that this could reveal further genotype-phenotype connections and potentially also suggest functional connections for known ones. In our study we focus on the genome marker dataset provided by the Wellcome Trust Case-Control Consortium (WTCCC), which includes SNP data for 14,000 disease samples and 3,000 controls [7].

References

[1] D. F. Easton and R. A. Eeles. Genome-wide association studies in cancer. *Human Molecular Genetics*, 17(R2):R109–15, 2008.

[2] R. A. King, J. I. Rotter, and A. G. Motulsky, eds. *The Genetic Basis of Common Diseases*. Oxford University Press, 2002.

[3] P. Kraft, Y. C. Yen, D. O. Stram, J. Morrison, and W. J. Gauderman. Exploiting gene-environment interaction to detect genetic associations. *Human Heredity*, 63(2):111–119, 2007.

[4] G. Lettre and J. D. Rioux. Autoimmune diseases: insights from genome-wide association studies. *Human Molecular Genetics*, 17(R2):R116–21, 2008.

[5] J. Marchini, P. Donnelly, and L. R. Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417, 2005.

[6] M. I. McCarthy and J. N. Hirschhorn. Genome-wide association studies: potential next steps on a genetic journey. *Human Molecular Genetics*, 17(R2):R156–65, 2008.

[7] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[8] A. Ziegler et al. Data mining, neural nets, trees – problems 2 and 3 of Genetic Analysis Workshop 15. *Genetic Epidemiology*, 31 Suppl 1:51–60, 2007.

34.7 Academic Activities

Membership in Program Committees

Kaustubh Patil:

– *3rd International Workshop on Practical Applications of Computational Biology and Bioinformatics*, Salamanca, Spain, June 2009.

34.7.1 Invited Talks

Alice McHardy:

– *Bioinformatische Verfahren zur Bewertung des Expressionsverhaltens von Metagenom-Kandidatengenen*, Invited talk, DECHEMA Society for Chemical Engineering and Biotechnology e.V., Frankfurt, March 2009.

- *Informatik für Metagenome: Einblicke in die Welt der unkultivierbaren Mikroben*, Workshop 'Wissenschaftsjournalismus: Schreiben über Informatik', IBFI Schloss Dagstuhl, September 2008.
- *Methods for the binning of metagenome sequence samples*, Workshop 'Statistical Methods in Metagenomics', Harnack-Haus, Berlin, July 2008.

Kaustubh Patil:

- *Application of data mining and machine learning tools to text mining and biosystems,* Tata Research Design and Development Centre, Pune, India, November 2008.

34.8 Teaching Activities

Winter Semester 2008/2009

Courses:

Bioinformatics for Metagenomics, in lecture series 'Introduction to Bioinformatics' (A. McHardy).

34.9 Dissertations, Habilitations, Offers, Awards

34.9.1 Offers for Faculty Positions

Ben Adams:

- University of Bath, 2008.

34.9.2 Awards

Alice McHardy:

- Sponsorship 'Fast Track Program' by Robert Bosch Foundation, July 2008.

34.10 Publications

Journal articles

[1] K. H. Gartemann, B. Abt, T. Bekel, A. Burger, J. Engemann, M. Flügel, L. Gaigalat, A. Goesmann, I. Gräfen, J. Kalinowski, O. Kaup, O. Kirchner, L. Krause, B. Linke, A. McHardy, F. Meyer, S. Pohle, C. Rückert, S. Schneiker, E. M. Zellermann, A. Pühler, R. Eichenlaub, O. Kaiser, and D. Bartels. The genome sequence of the tomato-pathogenic actinomycete clavibacter michiganensis subsp. michiganensis ncppb382 reveals a large island involved in pathogenicity. *The Journal of Bacteriology*, 190(6):2138–2149, 2008.

[2] L. Z. Holland, R. Albalat, K. Azumi, E. Benito-Gutiérrez, M. J. Blow, M. Bronner-Fraser, F. Brunet, T. Butts, S. Candiani, L. J. Dishaw, D. E. Ferrier, J. Garcia-Fernàndez, J. J. Gibson-Brown, C. Gissi, A. Godzik, F. Hallböök, D. Hirose, K. Hosomichi, T. Ikuta, H. Inoko, M. Kasahara, J. Kasamatsu, T. Kawashima, A. Kimura, M. Kobayashi, Z. Kozmik, K. Kubokawa,

V. Laudet, G. W. Litman, A. McHardy, D. Meulemans, M. Nonaka, R. P. Olinski, Z. Pancer, L. A. Pennacchio, M. Pestarino, J. P. Rast, I. Rigoutsos, M. Robinson-Rechavi, G. Roch, H. Saiga, Y. Sasakura, M. Satake, Y. Satou, M. Schubert, N. Sherwood, T. Shiina, N. Takatori, J. Tello, P. Vopalensky, S. Wada, A. Xu, Y. Ye, K. Yoshida, F. Yoshizaki, J. K. Yu, Q. Zhang, C. M. Zmasek, P. J. de Jong, K. Osoegawa, N. H. Putnam, D. S. Rokhsar, N. Satoh, and P. W. Holland. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Research*, 18(7):1100–1111, 2008.

[3] M. G. Kalyuzhnaya, A. Lapidus, N. Ivanova, A. C. Copeland, A. McHardy, E. Szeto, A. Salamov, I. V. Grigoriev, D. Suciu, S. R. Levine, V. M. Markowitz, I. Rigoutsos, S. G. Tringe, D. C. Bruce, P. M. Richardson, M. E. Lidstrom, and L. Chistoserdova. High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotechnology*, 26(9):1029–34, 2008.

[4] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6):495–500, 2007.

[5] A. McHardy, L. Krause, T. W. Nattkemper, A. Puhler, J. Stoye, and F. Meyer. GISMO - prokaryotic gene identification using a support vector machine for ORF classification. *Nucleic Acids Research*, 32(4):540–549, 2007.

[6] A. McHardy, H. G. Martin, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, 4(1):63–72, 2007.

[7] C. Meisinger-Henschel, M. Schmidt, S. Lukassen, B. Linke, L. Krause, S. Konietzny, A. Goesmann, P. Howley, P. Chaplin, M. Suter, and J. Hausmann. Genomic sequence of chorioallantois vaccinia virus Ankara, the ancestor of modified vaccinia virus Ankara. *Journal of General Virology*, 88(12):3249–3259, 2007.

[8] K. Patil and A. Kulkarni. A simple visualization technique to understand the system dynamics in bioreactors. *Biotechnology Progress*, 23(5):1101–1105, 2007.

[9] K. Patil, P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni. Ant colony optimization in the direct ordering gene expression data. *Journal of Hybrid Computing Research*, 1(1):10–10, 2008.

[10] S. Schneiker, O. Perlova, O. Kaiser, K. Gerth, A. Alici, M. Altmeyer, D. Bartels, T. Bekel, S. Beyer, E. Bode, H. Bode, C. Bolten, J. Choudhuri, S. Doss, Y. Elnakady, B. Frank, L. Gaigalat, A. Goesmann, C. Groeger, F. Gross, L. Jelsbak, L. Jelsbak, J. Kalinowski, C. Kegler, T. Knauber, S. Konietzny, M. Kopp, L. Krause, D. Krug, B. Linke, T. Mahmud, R. Martinez-Arias, A. McHardy, M. Merai, F. Meyer, S. Mormann, J. Mu?oz-Dorado, J. Perez, S. Pradella, S. Rachid, G. Raddatz, F. Rosenau, C. Rückert, F. Sasse, M. Scharfe, S. C. Schuster, G. Suen, A. Treuner-Lange, G. Velicer, F.-J. Vorhölter, K. J. Weissman, R. D. Welch, S. C. Wenzel, D. E. Whitworth, S. Wilhelm, C. Wittmann, H. Blöcker, A. Pühler, and R. Müller. Complete genome sequence of the myxobacterium Sorangium cellulosum. *Nature Biotechnology*, 27(11):281–1289, 2007.

[11] C. Schoen, J. Blom, H. Claus, A. Schramm-Glück, P. Brandt, T. Müller, A. Goesmann, B. Joseph, S. Konietzny, O. Kurzai, C. Schmitt, T. Friedrich, B. Linke, U. Vogel, and M. Frosch. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3473–3478, 2008.

[12] J. Vieira, C. Cardoso, J. Pinto, K. Patil, P. Brazdil, E. Cruz, C. Mascarenhas, R. Lacerda, A. Gartner, S. Almeida, H. Alves, and G. Porto. Location of a putative gene contributing to the setting of CD8+T lymphocytes: A modifier of hereditary hemochromatosis expression? *American Journal of Hematology*, 82:509–509, 2007.

[13] J. Vieira, C. Cardoso, J. Pinto, K. Patil, P. Brazdil, E. Cruz, C. Mascarenhas, R. Lacerda, A. Gartner, S. Almeida, H. Alves, and G. Porto. A putative gene located at the MHC-CLASS I region around the D6S105 marker contributes to the setting of CD8+ T lymphocyte numbers in humans. *International Journal of Immunogenetics*, 34(5):359–367, 2007.

[14] F. Warnecke, P. Luginbühl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S. G. Tringe, M. Podar, H. G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N. C. Kyrpides, E. G. Matson, E. A. Ottesen, X. Zhang, M. Hernández, C. Murillo, L. G. Acosta, I. Rigoutsos, G. Tamayo, B. D. Green, C. Chang, E. M. Rubin, E. J. Mathur, D. E. Robertson, P. Hugenholtz, and J. R. Leadbetter. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450(7169):560–565, 2007.

Book chapters and invited articles

[1] B. Adams, A. McHardy, C. Lundegaard, and T. Lengauer. Viral bioinformatics. In D. Frishman and A. Valencia, eds., *Modern Genome Annotation. The Biosapiens Network*, hardcover 8.1, pp. 429–452. Springer, New York, 2009.

[2] A. McHardy. Finding genes in genome sequence. In J. M. Keith, ed., *Bioinformatics. - Vol. 1, Data, Sequence Analysis and Evolution*, *Methods in Molecular Biology*, vol. 452, ch. 2, pp. 163–77. Humana Press, Clifton, N.J., USA, 2008.

[3] A. McHardy and B. Adams. Host evasion and the evolution of influenza. *PLOS Pathogens*, in preparation, 2009.

[4] A. McHardy and I. Rigoutsos. What's in the mix? Methods for the phylogenetic classification of metagenome sequence samples. *Current Opinion in Microbiology*, 10(5):499–503, 2007.

Various

[1] A. McHardy and K. Patil. Informatik für die Metagenomforschung: Einblicke in die Welt der unkultivierbaren Mikroorganismen, 2009.

[2] A. McHardy and L. Steinbrück. Method and system for building a phylogeny from genetic sequences and using the same for recommendation of vaccine strain candidates for the influenza virus, 2008.

Part IV

Index

Index