Eighth Biennial Report

April 2005 – March 2007

Contents

Part I

Overview – The Institute

1 Mission

Informatics: Accelerated progress in computing through new algorithms

While the acceleration of hardware has been a landmark of progress in computing technology in the past few decades, the computing enhancements that it provides is dwarfed by the increase in speed, performance, and robustness resulting from new algorithms. As a point in case, the status of hardware and algorithms in 1970 enabled to compute an optimal tour of a traveling salesman (a classical optimization problem and accepted benchmark for computing power) through 120 cities. Increasing the number of cities from n to $n+1$ leads to a multiplicative increase of the number of possible tours by a factor of n. Thus, relying only on the increase of hardware speed, with today's technology, and the algorithms of 1970 we could find optimal tours among only 135 cities. It is the progress in algorithms that, today, enables us to find optimal tours between many thousand of cities. Relying only on progress in hardware this performance would not be achievable in hundreds of years.

The Max Planck Institute for Informatics is devoted to cutting-edge research in informatics with a focus on algorithms and their applications in a broad sense. Our research ranges from foundations (algorithms and complexity, programming logics, automation of logic) to a variety of application domains (computer graphics, geometric computation, constraint solving, database systems and internet search, machine learning, and computational biology/bioinformatics).

- 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.

Our goal is to have impact through publications, software and people alike.

We regard informatics as a field that strives towards grounding the use of computational resources on thorough understanding of the underlying principles of computational methods. This involves reasoning mathematically and/or formally about the behavior of algorithms wherever possible. Consequently, much of our fundamental research is of a fairly rigorous mathematical character. Some of that research lies in established fields of theoretical computer science (algorithms and complexity, programming logics, automation of logic). Before a concrete application background, many computational problems are so complex that their in-depth formal treatment is not feasible. Therefore, in these cases, our analysis of the involved algorithms is more experimental, usually in the form of systematic validation based

on carefully curated application data and specially developed statistical models and, last but not least, of real life systems use in the application field. In fact, many problems in complex application areas are fuzzily stated or ill-formed, at first, such that modeling, i.e., giving a rigorous definition of a problem is a major aspect of the research. We test the value of our ideas by integrating new algorithms into systems and assessing their utility in realistic settings. This experience is useful in the short term in refining our designs and invaluable in the long term in advancing our knowledge. Most of the major advances in informatics and algorithms research have come through this combination of new theoretical insights and experimental validation. 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. We are also strongly committed to communicating our results. Exposing and testing our ideas in 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. We encourage our junior researchers to build their own research programs and to spread out over other institutions.

2 Overview

Summary

Since the last visit of the Scientific Advisory Board in June 2005, the institute has undergone a number of changes.

After the untimely passing of Harald Ganzinger, the department D2 on programming logics has been significantly downsized and will be phased out. For finding a new director, we followed the recommendation by the Scientific Advisory Board to conduct a broader search in various fields that are consistent with the mission of the institute. This strategy has been approved by the strategy board (Perspektivenkommission) of the CPTS (Chemical-Physical-Technical Section of the Max-Planck Society) in October 2005. We identified strategically relevant areas that pose long-term challenges and offer fruitful research opportunities within the field of Informatics. Subsequently we eliminated areas that would not provide sufficient synergies with the current areas of the institute or would significantly overlap with existing departments or the foreseen activities of the new MPI for Software Systems. This way we identified five areas of interest: Programming Logics, Machine Learning, Computer Vision, Robotics, Security and Complexity. We invited outstanding scientists from these areas for a distinguished colloquium series and drew on their expertise for identifying candidates, ideally in the age group around 40 years and possibly including the experts themselves. We have made good progress towards narrowing down on a small number of promising candidates, and we hope to initiate the next steps this year.

The leaders of our two independent research groups (IRGs), Fritz Eisenbrand and Marcus Magnor, accepted offers for full professorships at universities (Dortmund and Braunschweig). This has been an excellent step in their careers, and consequently the IRGs at the institute have been phased out at the end of 2006. A new IRG will be established at the institute in 2007, in the bioinformatics area, through the Max-Planck Society's highly competitive funding program for IRG leaders.

Two research groups (RGs), smaller units funded from the institute's internal budget, were established in September 2005 and January 2007, respectively. RG1 works on the automation of logic and is headed by Christoph Weidenbach, who also serves as the institute's research coordinator. RG2 works on machine learning and is headed by Tobias Scheffer.

The newly established Max-Planck Institute for Software Systems is operational since August 2005. It is partly located in Saarbrücken (and currently hosted in our institute building) and partly in Kaiserslautern, and it is headed by its founding director Peter Druschel. Our administration and its IT support group were restructured and have become shared services of both institutes. The joint administration continues to be headed by Volker Geiß; the reformed joint Information Services and Technology (IST) group is headed by Jörg Herrmann.

The International Max Planck Research School for Computer Science (IMPRS-CS), the institute's program for graduate education, was evaluated in October 2004 with very posi-

tive assessment, and our subsequent proposal for extending the school's extra funding by the Max-Planck Society until the year 2013 was approved in July 2005. The Max Planck Center for Visual Computing and Communication (MPC-VCC), the institute's program for independent junior scientists, was evaluated in November 2005 with very positive assessment; its funding extends until 2009, and new funding instruments beyond this point are being investigated.

Last but not least, the Scientific Advisory Board has been newly formed, with some of the previous members leaving and some new members joining. The leaving members are Mike Fischer (the board's previous chair), Manfred Broy (previous vice-chair), John Kececioglu, Mark Overmars, and Amir Pnueli. We are very grateful for the excellent advice that we obtained from the board. The new board will have the following members: Pankaj Agarwal, Douglas Brutlag, Joseph Hellerstein, Yannis Ioannidis, Friedhelm Meyer auf der Heide, Eugene Myers, Frank Pfenning, Claude Puech, Eva Tardos, and Demetri Terzopoulos.

Presently 116 doctoral students and 57 postdocs are affiliated with the institute. The scientific staff is complemented by the administration with 14.5 members (including secretaries), the IST group (6.5 members of staff), and our library (2 members of staff). 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 12.10, 13.9, 14.11, 15.15, 16.9, 17.11, 18.4, and 20.6. Within the reporting period 20 doctoral dissertations and 3 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 2005-2007, 25 IMPRS-CS students have successfully completed their PhD.

Members of the institute together with their colleagues at the CS department of Saarland University participate in the Graduiertenkolleg (Graduate Research Center) on Quality Guarantees for Computer Systems. The graduate research center comprises a program of collaborative research projects and advanced graduate courses on quality guarantees for computer systems with an emphasis on predictable runtime, provable correctness, and guaranteed quality of service. The program offers several PhD fellowships in close collaboration with the IMPRS-CS.

In 2007, we received 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.

In the year 2000 we have expanded the scope of our activities to also include bioinformatics. The CBI – Center for Bioinformatics Saar, a joint operation of Saarland University and several research institutes, among them our institute, has become internationally visible and ranked at top place (among five) in the half-time review by its funding agency DFG in 2003. The institute contributes heavily to the curricula offered in computational biology by CBI (Bachelor, Master, PhD) by taking up over a third of the teaching in core bioinformatics offered in Saarbrücken.

Third Party Funding

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 12.12, 13.11, 14.13, 15.17, 16.11, 17.12, 18.6, 19.5, and 20.8.

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 Max Planck Society 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 MPC-VCC was evaluated in November 2005 with very positive assessment; its funding extends until 2009, and new funding instruments beyond this point are being investigated. The center currently encompasses 12 independent research groups (6 at Stanford and 6 at MPII).

Honors and Awards

Several outstanding honors and awards were conferred on members of the institute. Kurt Mehlhorn was awarded the Doctor of Mathematics, honoris causa, from the Faculty of Mathematics at the University of Waterloo, Canada, in 2006. Gerhard Weikum was named Fellow of the Association for Computing Machinery in 2005.

Two scientists of the institute were conferred with the Otto-Hahn medal of the Max Planck society honoring outstanding dissertations: Niko Beerenwinkel (D3) in 2005 and Andrey Rybalchenko (D5) in 2006. Out of the 39 Otto-Hahn medal winners in 2006, Andrey Rybalchenko together with 3 colleagues got the offer to lead a junior research group at a

Max Planck Institute of his choice. Hendrik Lensch received the Eurographics Young Researcher Award 2005. Also in 2005, the Ackermann award of the European Association of Computer Science Logic (EACSL) was given to Konstantin Korovin (D2). Michael Goesele (D4) received the Eduard Martin Award in 2005 where Saarland university yearly recognizes the best junior researcher. The Heinz-Billing price 2006, an award given annually by the Heinz-Billing association for scientific computing inside the Max Planck society, was won by Rafal Mantiuk (D4) with his contribution "High Dynamic Range Imaging". Ingo Wald (D4) received the innovation price of the state bank of the Saarland in 2005. Christian Theobalt (D4) received in 2005 the ATI Fellowship Award, an award established by the Canadian company ATI to recognize outstanding achievements on the field of computer graphics. Christian also received the Eduard Martin Award 2006. Martin Theobald (D5) received the Dissertation Award of the German Computer Society's Chapter on Database and Information Systems in March 2007 and the "Honorable Mention Award" from the 2007 ACM SIGMOD Dissertation Awards.

Eric Berberich, Andreas Karrenbauer (both D1) received the Günter-Hotz medal in 2005 and Fabian Suchanek (D5) in 2006, an award of the computer science department at Saarland University for the best Master-level graduates. Martin Fuchs (VDI price 2005) and Jürgen Gall (SMI-DAGM graduation award) also received awards for their master theses. Julia Luxenburger (D5) was awarded the graduate price of the German Computer Society's Local Chapter in 2005.

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

Research Results

In Parts II and III of this report we describe in detail the research programs and results obtained in the period April 2005 through March 2007 for the five departments and four (independent) research groups of the institute. For an introduction into the research highlights of the groups, please refer to Part II of this report. The publication output of the institute during the reporting period comprises more than 600 articles in journals, books or proceedings of major refereed international conferences. All references that are coauthored by MPII staff members are marked with "•" in the reference lists throughout Parts II and III of this report.

In addition, many of the institute's results are available to the public through software systems such as the LEDA and CGAL libraries of efficient algorithms (D1), the SPASS theorem prover for first-order logic (RG1), several graphics toolkits (D4), the TopX and Minerva systems for information retrieval (D5), and bioinformatics software (D3), part of which is offered freely via the internet (ARBY protein structure prediction server, geno2pheno server for analyzing HIV drug resistance, NOXclass classifier identifying protein-protein interaction types) and part of which is distributed commercially (FlexX program suite for molecular docking). Our software is widely used and has been validated in a variety of applications. Again, details can be found in Parts II and III.

Cooperations

Various cooperations exist across the institute's departments and research groups and with the research groups at Saarland University.

- D1 and D5 jointly participate in the EU project DELIS and collaborate on efficient algorithms for information retrieval with joint publications in the premier conferences VLDB 2006 and SIGIR 2007.
- D1 and RG1 have started to build linear arithmetic solver technology into first-order reasoning.
- D1 and D4 have joint research interests in the areas of surface reconstruction and mesh processing.
- D2 and RG1 jointly participate in the German research center AVACS.
- D3 engages in dialog, consultation and exchange with the other departments of the institute on issues such as enumerating graphs (for fast scanning of compound databases, dialog with D1), geometric analysis of surfaces of proteins with D4, and analysis and reduced description of protein motions with D1 and D4.
- D3 is in the initial stage of conceptualizing an open science web concept for focussed biological and medical domains, especially viral infections with D5, and we started a dialog on general machine learning aspects with the recently established RG2.
- D3 and D4 have identified the analysis of molecular motions as a joint topic of interest.
- D5 and RG1 pursue joint interests in ontologies and knowledge management.
- D5 and RG2 have started collaborating on mining Web query and click logs.

The institute held a two-day retreat in May 2006 with about 30 senior researchers, for mutual information and to foster collaboration. It is planned to hold such a retreat on an annual basis.

All groups have strong relations to researchers at Saarland university.

- D1 cooperates with the group of R. Seidel (on computational geometry), M. Bläser (on complexity), M. Pinkall (computer linguistics), G. Smolka (constraint programming), and R. Wilhelm (program analysis).
- D2 cooperates with the group of P. Loos on collaborative business process modelling. D2 as well as RG1 cooperate in the context of the AVACS research center with H. Hermanns (stochastic processes), B. Finkbeiner (transition systems) and R. Wilhelm (shape analysis).
- D3 is a major partner in the Center of Bioinformatics Saar (CBI, `http://www.cbi-saar.de`), a joint initiative of Saarland University, MPII and the Fraunhofer Institute for Biomedical Engineering, Sankt Ingbert. In particular, there are cooperations with the group of S. Zeuzem (HCV), E. Heinzle (Metabolomics), and A. Meyerhans (HIV).
- D4 collaborates in the areas of realtime ray tracing and interactive global illumination (Ph. Slusallek), mesh processing and 3D video processing (J. Weickert), speech synchronization (M. Pinkal, H. Uszkoreit), and modeling and animation of synthetic virtual characters (W. Wahlster).

– D5 collaborates with researchers at Saarland University in the areas of data mining on gene expression data (H.-P. Lenhof), microstructure images in material sciences (F. Mücklich) computational linguistics for information extraction (H. Uszkoreit, M. Pinkal), data mining on software development histories (A. Zeller), anonymity and censorship resistance in peer-to-peer networks (M. Backes), and link analysis for impact assessment (R. Wilhelm).

Professional Activities

Members of the institute were involved in the organization of 21 workshops and conferences. In 186 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.

Offers of Faculty Positions

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

Surender Baswana (D1, IIT Kanpur)
Volker Blanz (D4, University of Siegen)
Friedrich Eisenbrand (IRG2, University of Dortmund)
Stefan Gumhold (D4, University of Dresden)
Vlastimil Havran (D4, University of Prague)
Ioannis Ivrissimtzis (D4, Durham University)
Jan Kautz (D4, University College London; Imperial College London (declined))
Marcus Magnor (IRG3, University of Braunschweig)
Ulrich Meyer (D1, University of Frankfurt)
Nabil Mustafa (D1, Lahore University of Management Sciences)
Michael Neff (D4, UC Davis, University of California)
Seth Pettie (D1, University of Michigan)
Andreas Podelski (D2, University of Freiburg)
Jörg Rahnenführer (D3, University of Dortmund; University of Kiel (declined))
Holger Theisel (D4, University of Bielefeld)
Nicola Wolpert (D1, University of Applied Sciences Stuttgart)

Part II

Overview – The Research Units

3 The Algorithms and Complexity Group (D1)

Kurt Mehlhorn's goals for D1 are threefold:

- do first-class basic research in theoretical and experimental algorithmics,
- build demonstrators, generally in the form of useful software libraries and implementations for specific application areas, and
- give junior researchers the possibility to work in a stimulating environment and to develop their own research programs and groups.

We are doing well in all three aspects. Our research group has impact through our publications and software projects, and through our researchers, most of whom move on to jobs in academia and industry. We conduct fundamental research on a broad range of topics. Our work appears in top conferences and journals. Our demonstrators, most notably the CompleteSearch Engine and software libraries CGAL and EXACUS, have significant impact in academia, as experimental platforms, and in industry, as a means to transfer knowledge. Kurt Mehlhorn is the only permanent member of the group. Other researchers stay in the group between one and seven years. More than half of the current group members have joined the group since the last review. Conversely, more than half of the group at the time of the last review have moved on to positions in academia and industry. We also have an extensive short- and long-term visitor program. Collaboration with other groups at the institute has increased. We cooperate intensively with Gerhard Weikum's group in information retrieval, and we cooperated with Fritz Eisenbrand's group in theoretical and applied optimization.

The effort expended for items one and two is in a ratio of approximately three to one. Many group members contribute in both areas, and the items complement each other. Much of our experimental work is based on our theoretical insights. For example EXACUS profits from our improved algorithms for root isolation of univariate polynomials and topology determination of algebraic curves. The CompleteSearch Engine relies on a new index data structure that is more powerful than standard index structures for large data sets and yet requires no more space. Our work on shortest path computations in road maps is inspired by theoretical work on graph spanners. Conversely, the experimental work has suggested new theorems – e.g., the average case analysis of matching algorithms and the mathematical analysis of latent indexing – and sharply points to deficiencies in the theory. Moreover, it is highly rewarding to see how our demonstrators and libraries are used by other researchers and how they lead to industrial applications.

We are currently pursuing six research areas, each having its own coordinator(s). The subgroups and their coordinators are:

- *Foundations and Discrete Mathematics, coordinator: Benjamin Doerr.* Seth Pettie coordinated the Foundations group until September 2006; he is now an Assistant Professor at the University of Michigan. Benjamin Doerr joined the group from the University of Kiel in the Spring of 2005 and increased the discrete math activity of the group.

- *Combinatorial Optimization, coordinators: Ernst Althaus and Kurt Mehlhorn.* Ernst Althaus returned from a three year stay at INRIA LORIA.

- *Computational Geometry, coordinators: Stefan Funke and Joachim Giesen.* Stefan Funke and Joachim Giesen are both funded under the Stanford-MPG center. Stefan returned from Stanford and Joachim joined from ETH Zürich.

- *Advanced Models of Computation, coordinator: Uli Meyer.* Uli Meyer was hired as a full professor at Frankfurt University in the Spring of 2007. We will have to find a new coordinator.

- *Information Retrieval, coordinator: Holger Bast.* Holger Bast built this group up over the past three and a half years. It is now running full steam and the first PhD candidates have submitted their theses.

- *Geometric Computing, coordinator: Michael Sagraloff.* Lutz Kettner coordinated this group until Spring 2006; he is now with Mental Images GmbH. Michael Sagraloff was previously in the Math Department of the Universität des Saarlandes.

In the following sections, I discuss (1) our research directions and representative achievements, (2) our research strategy, (3) software development, (4) group development, (5) cooperations and (6) industrial interactions.

Research Areas and Representative Achievements

We are currently pursuing six research areas. For each area I highlight some results and explain how the group operates.

Foundations and Discrete Mathematics, Coordinator Benjamin Doerr: Randomized rounding is a versatile tool for obtaining integral solutions to optimization problems from non-integral solutions. Non-integral solutions are frequently easier to obtain. Of course, the rounding usually destroys optimality (but frequently retains approximate optimality); it may also destroy feasibility, e.g., if there is a constraint fixing the sum of the variables. B. Doerr has shown how to extend randomized rounding so that it can cope with hard side constraints. The work is part of a general effort in discrepancy theory that studies approximations of something large, complicated or continuous by something small, simple or discrete, while preserving certain properties, see Section 12.3.1.

Evolutionary algorithms, such as simulated annealing and ant algorithms, have always been popular among practitioners because they are easy to implement and widely applicable. The algorithms community has only recently started to investigate evolutionary algorithms

from a theoretical perspective. F. Neumann (in joint work with C. Witt) gave the first rigorous runtime analysis of an ant colony algorithm for a combinatorial optimization problem, see Section 12.3.2.

Computer algebra, see Section 12.3.3, became a topic of the group because of its necessary applications to our applied work in computational geometry. A. Eigenwillig (in joint work with V. Sharma and Ch. Yap) improved the analysis of Descartes method for real root isolation and also showed how the method can cope with multiple roots.

In the area of game theory and mechanism design, see Section 12.3.4, G. Christodoulou and A. Kovacs obtained improved upper and lower bounds for scheduling on unrelated machines.

Graph algorithms and Data Structures, see Sections 12.3.6 and 12.3.5, are wide areas which has always been a focus of our work. S. Baswana and T. Kavitha obtained improved algorithms for computing spanners and approximate distance oracles 12.3.6.

S. Pettie improved the analysis of pairing heaps – a practical and efficient data structure introduced 20 years ago by Fredman, Sedfewick, Sleator, and Tarjan, whose basic complexity remained open. He gave the first sub-logarithmic time bound on decreasekey. See Section 12.3.6.

The Foundations and Discrete Math group pursues a multitude of research themes and we consider this one of our strengths. The steady influx of new researchers and new ideas rejuvenates our concept of what constitutes an important, foundational problem in computer science. Therefore, it is impossible to precisely forecast the course of our future research. However, in general terms, our goals can be simply stated. We will pursue optimal algorithms for the classical (and most stubborn) problems in computer science and attempt to explain the inherent difficulty of these problems. Likely topics include graph theory, combinatorial optimization, algebra, and data structures. We will study abstracted versions of problems encountered in real-world applications and explain the empirical behavior of algorithms using rigorous theoretical models. Although the immediate consequences of our work are largely theoretical, we expect that a byproduct of our research will be simple and effective algorithms fit for widespread use.

Combinatorial Optimization, Coordinators Ernst Althaus and Kurt Mehlhorn Mehlhorn: R. Beier (jointly with H. Roeglin and B. Voecking) extended his work on smoothed analysis of algorithms for NP-complete problems, see Section 12.4.1. Approximation algorithms, see Sections 12.4.2 and 12.4.3 are pursued within the group and in our partner group headed by Naveen Garg at IIT Dehli. N. Garg and A. Kumar obtained improved results for minimizing flow time on related machines. In this problem, jobs arrive in the system and have to be scheduled on machines with different speeds. The flow time of a job is the time between its arrival and its completion and the goal is to minimize total flow time.

We also do much work in the area of applied optimization. The trunk packing project, see Section 12.4.4, aims at software that can pack a car trunk according to the European DIN and the American SAE norm. The project is in collaboration with Daimler-Chrysler and has moved to Mainz with E. Schömer. The solution for the DIN norm, developed by F. Eisenbrand, S. Funke, A. Karrenbauer, J. Reichel, and E. Schömer, is now in use at Daimler-Chrysler. Algorithms for computing shortest paths in road networks have received much

attention recently. H. Bast, S. Funke, and D. Matijevic (in collaboration with P. Sanders and D. Schultes, now at Karlsruhe) have developed a new approach that answers queries two orders of magnitudes faster than any previous solution, see Section 12.4.4. The approach was inspired by the work on graph spanners. E. Althaus and R. Naujoks developed a new exact algorithm for computing minimum cost Steiner trees for a set of strings under the Hamming distance, see Section 12.4.5. The algorithm has applications to computing evolutionary trees and can solve practical problem instances that were outside the reach of previous approaches. The improved running time results from a deeper theoretical understanding of the problem.

In optimization, we cooperated closely with Fritz Eisenbrand's group.

Computational Geometry, Coordinators: Stefan Funke and Joachim Giesen: We worked along three axes: sample-based geometry (Section 12.5.1, geometry in wireless networks (Section 12.5.2, and fundamental data structures and algorithms (Section). In wireless networks the possibility to communicate is determined by geometric proximity. Thus the connectivity of the network encodes geometric information about the distribution of the nodes in space. Can this information be retrieved? S. Funke with various co-workers (Ch .Klein, N. Milosavljevic, L. Guibas, A. Nguyen, Y. Wang, S. Laue, R. Naujoks) obtained a number of results in this direction. For example, they showed that holes (large areas void of nodes) can be detected, that a sketch of the topology of the network can be obtained, and that routing protocols can be made more robust by exploiting this information.

A. Eigenwillig, L. Kettner, and N. Wolpert extended snap rounding from the linear domain to Bezier curves. They show how to snap round (control points are moved to lattice points) an arrangement of Bezier curves without destroying the topology, i.e., elements may be identified but do not move across each other, and without prior computation of the exact arrangement. N. Mustafa and S. Ray showed that weak ϵ-nets have a basis of size $O(1/\epsilon \log(1/\epsilon))$ in any dimension. K. Elbassioni, A. Fishkin, N. Mustafa, and R. Sitters obtained approximation algorithms for variants of the Euclidean Traveling Salesman Problem.

Advanced Models of Computation, Coordinator Uli Meyer: Graph Exploration is a notoriously hard problem for cache-efficient and external memory algorithms. U. Meyer, D. Ajwani, V. Osipov, and R. Dementiev (now Karlsruhe) performed a detailed experimental study of various proposals for external BFS and added BFS to the library STXXL of external memory algorithms, see Section 12.6.1 S. Govindra and co-workers obtained an improved algorithm for well-separated pair decomposition, 12.6.1. Our activities in streaming algorithms 12.6.2 and lock-free data structures 12.6.3 are new.

Lars Arge (Univ. of Aarhus) succeeded in attracting a center for external memory computation funded by the Danish government. External partners are MPI (Mehlhorn), MIT (Demaine and Indyk), and Frankfurt (Meyer).

Information Retrieval, Coordinator Holger Bast: Holger Bast built this group over the past three and a half years. The goal of the group is to make intelligent search fast. The work of the group spans from theoretical work on new index structures to building a demonstrator, the CompleteSearch Engine, see Section 12.7.1. H. Bast and D. Majumdar gave the first satisfactory theoretical explanation of the behavior of spectral methods in information

retrieval. Spectral retrieval methods are used to uncover the semantic concepts that underlie a document collection, by means of an eigenvalue analysis of the so-called word-document matrix. Bast and Majumdar observed that spectral retrieval methods work by assigning "relatedness" scores to pairs of words, and they identified the essential word co-occurrence patterns that lead to high "relatedness" scores. Another highly interesting result is an efficient search engine with auto-completion. Bast and co-workers engineered an expert search engine with a novel context-sensitive autocompletion feature. At its core is a sophisticated new indexing data structure that enables extremely fast response times even in the worst-case and is also space-optimal. A number of data sets are indexed by this search engine, e.g., the MPI web-site, Wikipedia, and the databases of service requests to our computer support group and our library. The auto-completion mechanism can also be used to implement faceted search.

Geometric Computing, Coordinator: Michael Sagraloff: Computational geometry traditionally studied simple objects (points, lines, planes, spheres) and avoided higher-degree algebraic objects (i.e., objects defined by higher degree algebraic equations). These objects were studied by the Solid Modeling community. Only recently, the computational geometry community started to focus on curved objects (our group was among the first to do so), while maintaining the goal of exact computations.

At the last review, we announced the first release of the EXACUS library (Berberich, Eigenwillig, Hemmer, Kettner, Kerber, Sagraloff, Schömer, and Wolpert). It supports the exact, complete and efficient computation of arrangements of low degree algebraic curves (arbitrary degree-two and three curves and a subclass of degree-four curves). In the meantime, the CGAL consortium has adopted the EXACUS approach and EXACUS is being integrated into CGAL. We also extended the functionality (rotated conics, lower envelopes of quadrics). A. Eigenwillig, M. Kerber, and N. Wolpert designed a new algorithm for analyzing a single algebraic curve. The algorithm makes use of the bitstream Descartes method obtained in the last period.

Controlled perturbation is an alternative approach to robust geometric computation. The idea is to solve the problem at hand on a perturbed input. The hope is that the perturbation helps to avoid degenerate situations and reduces the demand in terms of high-precision arithmetic. K. Mehlhorn, R. Osbild, and M.Sagraloff developed a general technique for analyzing perturbation approaches.

The coordinators coordinate research and teaching in their area and together with Kurt Mehlhorn and the senior researchers form the steering committee of D1. The steering committee meets at least once monthly to discuss and organize the work of the group. We use our noon seminar, reading group meetings, mini courses, and group meetings to educate each other and to inform other group members about own work. In the fall of each year, we have one day retreat at which senior group members present their research programs and all group members give short talks about their work.

I encourage group members to build their own subgroups and to apply for their own research funding, which they can take with them once they leave the group. At the moment, H. Bast, E. Althaus, and U. Meyer have their own DFG-funded projects and J. Giesen

and S. Funke have support from the Stanford-MPG program. The group participates in EU-projects ACS and DELIS and in a GIF-project.

Research Strategy

How do we choose our research areas and our specific research questions?

Our choice of research areas is long-term. The areas foundations, combinatorial optimization, models of computation, computational geometry and software libraries have existed since the founding of D1 in 1990 and are core areas of algorithms. Except for software libraries, they can be found in almost any algorithms group. The fact that we also build (successful) software is a distinctive feature of our group. I have made sure over the past 15 years, that at least one senior researcher, some postdocs and some PhD-students work in each of our core areas. For example, software libraries was coordinated by Stefan Näher, Stefan Schirra, Lutz Kettner and now Michael Sagraloff; models of computation was coordinated by Michael Kaufmann, Jop Sibeyn, Peter Sanders, and now Uli Meyer; foundations was coordinated by Torben Hagerup, Susanne Albers, Berthold Vöcking, Seth Pettie, and now Benjamin Doerr. Of course, the specific research questions change over time, e.g., within foundations there are now strong efforts on discrepancy theory (one of Benjamin's strengths) and evolutionary algorithms (which came to the group via postdoc Frank Neumann and has caught the interest of other group members).

Other areas come and go and are usually associated with a person. In most cases, the person joined the group as a junior researcher, suggested a new area and I gave him/her the resources to develop it: Petra Mutzel built up graph drawing and took the area with her to Vienna; Hans-Peter Lenhof built up Computational Biology and took the area with him to the University (of course, with Thomas Lengauer's group the area is now much more strongly represented at the institute); Elmar Schömer built up Assembly and Simulation and took the area with him to Mainz; Holger Bast has built up Information Retrieval; Stefan Funke is building an effort at the interface of geometry and networking; and Joachim Giesen is setting up a group at the interface of geometry and learning. In either case, the crucial task is to identify the right persons and to provide the right environment.

How do we choose our specific research questions? There are two main strategies:

- By hiring excellent postdocs, who bring new problems into the group, and by fostering an atmosphere of cooperation. Our noon seminars, the mini-courses and the integration-talks in the fall of each year are our tools for educating each other about new results and topics and for integrating new group members.

- Identifying long-term goals and building subgroups to work on them. The work on exact algorithms for curves and surfaces and the work on information retrieval belongs to this category. Frequently, we attract additional outside funds to support the work (here ECG/ACS and DELIS, respectively) and we install a teaching program to bring in master and PhD-students.

Software Projects

A distinctive feature of our group is the combination of theoretical and experimental research and software development. Our software serves the academic and research community as an experimental platform and a teaching aid and transfers knowledge to industry. We started with LEDA (Library of Efficient Data Types and Algorithms), then came CGAL (Computational Geometry Algorithm's Library) and BALL (Biochemical Algorithm's Library; it moved to Hans-Peter Lenhof's group at the University and Oliver Kohlbacher's group at Tübingen), EXACUS and STXXL (standard template library for external memory; it moved to Peter Sanders' group in Karlsruhe). The older libraries have achieved international recognition. The combined number of installations is several thousand and they are marketed through Algorithmic Solutions and Geometry Factory, respectively.

We also develop software for specific applications, e.g., in particular geometric packing (in a cooperation with Daimler-Chrysler) and fast intelligent search in large data sets (the CompleteSearch Engine).

Our software work is intimately tied to our theoretical work and inspirations flow in *both* directions. New algorithms lead to improved implementations and shortcomings of our implementations suggest new theoretical research.

Group Development

The composition of the group has changed considerably over the past two years.

On the senior level Uli Meyer (now full professor at Uni Frankfurt) and Lutz Kettner (now head geometric computing at Mental Images, Berlin) left the group. Ernst Althaus returned after a two year stay at INRIA Nancy. Stefan Funke returned after a year at Stanford. Joachim Giesen joined the group from ETH Zürich.

On the postdoc level, we had the expected fluctuation: Every year about eight new postdocs join the group. Integrating them into the group is a major task. Over the past two years, we attracted new postdocs from Yale, Chalmers, Patras, Duke, MIT, and Kiel. Many former postdocs moved on to academical jobs. Nicola Wolpert moved to a professorship at the University for Applied Sciences (Fachhochschule) Stuttgart. Seth Pettie became assistant professor at the University of Michigan. Nabil Mustafa became assistant professor at Lahore University of Management Sciences. Surender Baswana became an assistant professor at IIT Kanpur. Gianni Franceschini went to the University of Pisa. Lukasz Kowalik and Marcin Mucha went to the University of Warsaw. René Sitters went to Eindhoven University. Katarzyna Paluch went to the University of Wroclaw. Aleksei Fishkin now works at Siemens. Sadly, Martin Kutz passed away in January 2007.

Ten PhD-students have finished since the last review: Roman Dementiev (now research associate at Uni Karlsruhe), Joachim Reichel (now research associate at Uni Dortmund), Dimitris Michail (now postdoc at MPI), Irit Katriel (now postdoc at Brown University), Peter Hachenberger (now postdoc at Technical University Eindhoven), Annamaria Kovacs (now Uni Frankfurt), Debapriyo Majumdar (now IBM Research, Bangalore, India), Ingmar Weber (now postdoc at EPFL, Lausanne, Switzerland), Dieter Mitsche (now postdoc at UPC, Barcelona), and Domagoi Matijevic.

Ernst Althaus, Benjamin Doerr, Joachim Giesen, and Lutz Kettner completed the Habilitation. Holger Bast and Stefan Funke have submitted their Habilitation requests.

Cooperations

We list some long-term cooperations (short-term ones for single research papers are not included).

Local Cooperations Within the institute, we cooperate intensively with the groups of Gerhard Weikum and Fritz Eisenbrand (Fritz moved to Dortmund in 2006). On campus, our partners are Raimund Seidel, Markus Bläser, Michael Backes, the constraint programmers (Gert Smolka's group) and the computer linguists (Manfred Pinkal's group). In all cases, the collaborations led to joint publications.

Projects with External Funding We are part of the EU-Projects *ACS* and *DELIS*. The partners include INRIA Sophia Antipolis, ETH, FU Berlin, Tel Aviv University, La Sapienza Rome, Groningen, Patras, and Athens. The cooperation with these partners has a long tradition.

In CGAL, we continue to cooperate with Utrecht, INRIA Sophia-Antipolis, Berlin, Zürich, Tel Aviv, and Geometry Factory.

We cooperate with Uri Zwick in Tel Aviv under a grant from GIF (German-Israel-Foundation).

Industrial Interactions

Algorithmic Solutions (AS) is a spin-off of D1. AS is marketing LEDA (Library of Efficient Algorithms) under license agreements with the Max Planck Society. It is also responsible for maintenance of LEDA. CGAL is marketed by Geometry Factory, a spin-off of our EU-projects CGAL and GALIA.

LEDA is widely used in academia and industry. CGAL is also widely used in academia and has found some commercial use.

Our work on geometric packing is supported by Daimler-Chrysler.

4 The Programming Logics Group (D2)

Since the death of Harald Ganzinger and the subsequent move of Andreas Podelski to the University of Freiburg the number of members of the programming logics group has decreased a lot. During the last two years several former members have taken positions at other prestigious academic and research institutions, several others have moved to the research group (RG1) of Christoph Weidenbach. The work of Viorica Sofronie-Stokkermans, Uwe Waldmann, Carsten Ihlemann and Thomas Hillenbrand is reported in detail in Section 17.

Vision

"Software to Make Software Better"

The goal is to achieve predictable quality in software, just as for any other industrially engineered product. One way towards that goal is to integrate more and more automation into the software process (specification, programming, testing). This is true of every process: the more automated it is, the more reliable it gets.

The automation must be performed by software that manipulates software. That is, by algorithms and tools that take code as input (source-code or design models) and analyze the code as it is being written (and not only once the system is up running, usually as part of bigger system). They analyze the code and automate the bug finding, the testing and the quality-assurance procedures.

As utopical this may sound to an outsider, in a preliminary form these tasks are performed routinely already today, namely by static analysis and type checking in every modern compiler. What will lift static analysis to software automation in our sense is the integration of powerful constraint solvers, decision procedures and theorem provers. It is this integration with reasoning capabilities that turns tools into intelligent mathematical assistants in the software process: assistants that are able to mathematically reason about (what will happen during) execution.

Predictable quality and reliability go hand in hand with automation to serve another goal, which is to increase software productivity. Hence we would like to categorize our research under Software Productivity Tools.

Overview and Research Directions

The programming logics group is one of the two groups that exists since funding the Max Planck Institute for Informatics in 1991. Until his death in June 2004 the group was headed

by Harald Ganzinger. Thomas Lengauer took over as Acting Director. Main research topics of the programming logics group are automated theorem proving and deduction-based verification and analysis of complex systems and programs.

The emphasis of our vision and of our research efforts is shifting from a broad to a more concrete application perspective. This is a consequence of our participation in four related research projects: AVACS, Verisoft, eJustice and R4eGov.

Aim of deduction-based analysis is to provide algorithms and tools for the automatic verification of programs and systems properties, such as freeness of deadlocks and run time errors. A typical run time error can be a buffer overflow. In case of systems we investigate stability properties of hybrid systems as well as safety properties of public business processes. Hybrid systems are an example of such models, in this case of embedded systems that control some physical environment using sensors and actuators. The environment may show non-linear behavior (speed, temperature, ...). Here, a new approach for solving non-linear constraints has lead to a rather surprising decidability result for the correctness of 'robust' hybrid systems; arguably all practically interesting examples fall into this class of hybrid systems (robustness accounts for measuring errors). A new model checking tool for hybrid systems implements the corresponding decision algorithm, based on the new solver for non-linear constraints.

A crucial quality feature of applications and products is to guarantee absence of errors. In collaboration with leading research groups all over the world and several industrial partners we develop new algorithms and tools that can be integrated and combined into known techniques for classical program and system analysis, as they are used daily in computers today. This leads to increasing both the expressiveness and the efficiency of such programs. In the realm of decision procedures and constraint solving, this has lead to a new framework for the flexible and efficient integration of decision procedures into general-purpose SAT-checkers. For superposition theorem proving, we were able to show that switching from usual first-order logic to a logic with total and partial functions results in a calculus for which hierarchic and modular combinations are complete.

In the research direction of Software Model Checking, a sequence of results has lead to a breakthrough in automatic program verification. We are now able to check software for the full-fledged class of *temporal-logic* properties. This is the class of properties for which hardware is checked with great success today. The sequence of results goes from logic (a new proof rule) to algorithms (a new way of abstracting programs) and finally to tools, co-developed and evaluated in an industrial context (on Windows device drivers).

One of our research branches focuses on formal analysis of business processes aiming at ensuring security and interoperability. In terms of security we prove the correctness of access control algorithms, in particular the correct delegation and revocation of rights for accessing processes, data and documents. Formal modeling techniques, such as Petri nets are used to model and verify role based access control algorithms. Moreover, we use formal analysis for checking technical, semantic and organizational interoperability by proving correctness and controllability of executable processes and web services.

Projects and Cooperation

In the Transregio-SFB project AVACS (funded by the DFG) we collaborate with the universities of Oldenburg (Damm, Olderog), Freiburg (Becker, Nebel) and Saarbrücken (Wilhelm) as well as with MPI's RG1 (Christoph Weidenbach) on foundational research in tools for the automatic verification of safety-critical systems in planes, cars and trains. In the R4eGov project (IP project funded by the EU) we collaborate with SAP (France), Unisys (Belgium), University of Koblenz (Wimmer) and the DFKI Saarbrücken on the verification of security aspects in business processes and processes of public administrations in particular.

5 The Computational Biology and Applied Algorithmics Group (D3)

History of the Group

The department became established when Thomas Lengauer joined the institute on October 1, 2001. The department has grown somewhat in size throughout the reporting period. As this text is written there are 22 scientists (5 postdoc, 17 predoc), four predoc fellowship holders, two support people and three guests in the department, see `http://www.mpi-sb.mpg.de/units/ag3/people.html`. Four people left the department during the reporting period, six people joined the department during that time.

Due to the fact that the curricula on bioinformatics at Saarland University have gained momentum, there is a steady influx of students into the department who work on bachelor (FoPra), diploma and master theses in the department. Aside from the director, several scientists have offered in-depth courses on bioinformatics themes within the bioinformatics curricula at Saarland University.

Research focusses

In the reporting period, the group has expanded along existing focuses and created new ones. The main driving theme of *Bioinformatics and Computational Systems Biology for Diseases* has gained additional contours.

We implement this theme before the background of specific diseases, notably viral infections and malignant tumors. With regard to viral infections, we focus on AIDS, caused by HIV (see Section 14.4)and Hepatitis C, caused by HCV (see Section 14.6.3). On AIDS we have expanded beyond the analysis of viral resistance patterns, and are now also studying viral tropism (the choice by the virus of one of different paths for entering the host cell) and interactions between HIV and the human immune system (by analyzing MHC binding of HIV epitopes). Our work on HIV resistance analysis has progressed considerably and is now using data that are collected on a European scale. Six people in the department have contributed significantly to the research on HIV. Thus the personnel base for this topic has expanded further.

Our work on HCV has concentrated on helping to elucidate the role of viral proteins in the infection. Here we are continuing to put together process pipelines for bioinformatics analysis that are composed of our own tools and also of tools offered by other people. Our work has progressed beyond structural modeling of the proteins in question to the analysis of relevant protein interaction networks. For this purpose, we have developed several tools for visualization and analysis (Section 14.6). Mario Albrecht, who is the coordinator of most of the activities on HCV, has received his PhD and, in November 2006, has assumed the position of a group leader in the department. His group on "Molecular networks in Medical

Bioinformatics" is the focus of the research on this very dynamic field in computational biology. Five scientists are currently working on this topic. The topic of malignant tumors has been approached from the viewpoint of analyzing comparative genomic hybridization (CGH) data. On the basis of bioinformatical analysis of CGH data we could sharpen previous prognostic cancer classifications substantially. We used the "genetic progression score" (GPS) for this purpose (see Section 14.7.2). In the report period, we have progressed to analyzing arrayCGH data, the microarray variant of CGH data. Issues of determining statistically significant regions of constant copy number become important here. We also have expanded our work on cancer by starting a new research activity on epigenetics (see Section 14.7.4) In the first two years of this activity focused on basic research. Now we will embark on research that is focused on analyzing epigenetic foundations of cancer.

The more classical themes of the group, "Analysis of Protein Structure and Function" (see Section 14.5) and "Molecular docking" (see Section 14.8) are continuing to draw solid contributions and reflect the global progress in their respective fields.

The Group 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, MPII and the Fraunhofer Institute for Biomedical Engineering, Sankt Ingbert, which has received substantial funding from DFG.

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.

Some Achievements

Among the scientific achievements within the report period are:

- *HIV Bioinformatics* This project has gained much interest in the scientific community and beyond. We could place an invited review in the journal Nat Reviews Microbiology in October 2006 [6]. The project has contributed significantly to the European STREP *Euresist* in which resistance data are collected on a European scale. The project has gained much attention in the media, including a trailer on the national news.

- *Computational Epigenetics* This new area in the department, has already gained several major achievements. The first software development in the area, the BiQAnalyzer tool for low-level analysis of methylation data, has drawn several hundred users and is officially being recommended by one of the major suppliers of the experimental technology. Also, we could publish one of the first genome-wide analyses of associations between DNA sequence and methylation patterns [4].

- *Analysis of CGH data* The GPS score which has been described above was applied to refining the prognosis for prostate cancer which is traditionally expressed by the

Gleason score, a histological classification. The GPS score was so helpful in resolving the uncertain Gleason score class 7 that the work drew an award from the urology community in late 2005 (see 14.12.4).

Thomas Lengauer is the Conference Chair of of ISMB/ECCB 2007, the largest Bioinformatics Conference in the year 2007 which will be held in Vienna, Austria, in July 2007.

Honorary Memberships and Awards

Jochen Maydt, Oliver Sander and Tobias Sing have received awards for their outstanding diploma theses. Niko Beerenwinkel has received the Third-Place Heinz-Billing-Award 2004 of Max Planck Society for the software on resistance analysis of HIV. Bernd Wullich has earned the Wolfgang-Hepp-Prize 2005 for a project in collaboration with Jörg Rahnenführer. Christoph Hartmann, Oliver Mueller, and Oliver Sander have received Best Poster Awards on Conferences in 2006 and 2007 (see 14.12.4).

Third Party Funding

Currently 12 of the 26 scientists in the group are funded by third party money, 4 from the ministry of science and 4 from DFG, 2 from EU. Two fellowship holders are financed by the International Max Planck Research School for Computer Science.

Collaboration and Networking

Cooperation with the other departments in the Institute is progressing on the topics of *Surface Analysis of Protein Binding Sites* and *Motion Clustering*.

For a lab in Computational Biology, being placed within a computer science context requires special efforts for networking to the biology community. On campus the major device for networking to biologists is CBI. Our work on epigenetics, metabolomics and, in some cases, chemical biology, is in cooperation with CBI partners. The DFG Clinical Research Group on HCV is our major vehicle for cooperating with Medical Experts in Saarland University. However, our cooperation extends beyond the limits of this group to geneticists interested in cancer. On a National scale the Arevir consortium is continuing to stay together beyond the initial funding period. This consortium extends its research on HIV resistance and puts on an annual meeting for practicing doctors in Germany that informs them about the clinical use of the tools and progress in research. The *Euresist* project, in which we are a partner, is bringing HIV resistance research to a European level. Our work on the analysis structure-function relationships in proteins is largely carried out within the European BioSapiens Network of Excellence.

An alphabetical list of academic cooperation partners that are external to the institute follows:

Dr. Ahlenstiel, (NIH Bethesda) – Bioinformatics for HIV

Dr. Apostolakis (Bioinformatics, University of Munich) – Cheminformatics

Prof. Bartenschlager (Virology, University of Heidelberg) – HCV

Prof. Bernhardt (Biochemistry, Saarland University) – Docking and Homology Modeling

Dr. Beerenwinkel (Harvard University) – Bioinformatics for HIV
Dr. Cline (Institut Pasteur, Paris) – Protein Networks
Prof. Eils (German Cancer Research Center, Heidelberg) – Expression Data Analysis
Prof. Fessel (Medical Care Program Northern California) – Bioinformatics for HIV
Dr. Finn (Sanger Institute, Cambridge) – Domain Networks
Prof. Hartmann (Pharmacology, Saarland University) – Docking and Homology Modeling
Prof. Heinzle (Biotechnology, Saarland University) – Metabolomics
Dr. Heckmann-Pohl (Biotechnology, Saarland University) – Docking
Dr. Hermjakob (European Bioinformatics Institute, Cambridge) – DAS Protocol
Prof. Hovig (Tumor Biology, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Center, Norway) – Epigenetics
Dr. Kaiser (Virology, University of Cologne) – Bioinformatics for HIV
Dr. Krobitsch (MPI for Molecular Genetics, Berlin) – Neurodegenerative Diseases
Prof. Lenhof (Bioinformatics, Saarland University) – Bioinformatics Software
Prof. Mansmann (University of Munich) – Expression Data Analysis
Prof. Marian (Theoretical Chemistry, University of Düsseldorf) – Cheminformatics
Prof. Meyerhans (Virology, Medical Campus, Saarland University) – HIV Recombination, HCV
Prof. Middaugh (Pharmaceutical Chemistry, University of Kansas) – Protein Networks
Prof. Mutzel (Computer Science, University of Dortmund) – Graph Layout
Dr. Prlic (Sanger Institute, Cambridge) – DAS Protocol
Dr. Schönbach (RIKEN Genomic Sciences Center) – Analysis of DNA Sequences
Prof. Schreiber (Clinical Molecular Biology, University of Kiel) – Inflammatory Diseases
Prof. Schulz (Urology, University of Düsseldorf) – Expression Data Analysis
Prof. Shafer (Devision of Infectious Diseases, Stanford University) – Bioinformatics for HIV
Dr. Spengler (Medical Clinic and Policlinic I, University of Bonn) – Bioinformatics for HIV
Prof. Sonnerborg (Karolinska Institute) – Bioinformatics for HIV
Dr. Takken (Plant Pathology, University of Amsterdam) – Plant Immunity
Prof. Tosatto (University of Padova) – Protein Structure Prediction
Dr. Urbschat (Human Genetics, Medical Faculty, Saarland University) Glioblastoma
Prof. Valencia (National Cancer Research Center, Madrid) – Protein Networks
Dr. Waha (Department of Neuropathology, University of Bonn) – Epigenetics
Prof. Walter (Genetics/Epigenetics, Saarland University) – Epigenetics
Prof. Weickert (Computer Science, Saarland University) – Protein Structure
Prof. Wenz (Chemistry, Saarland University) – Cheminformatics
Prof. Wullich, (Urology, Medical Campus, Saarland University) – Bioinformatics for Prostate Cancer
Prof. Zang (Human Genetics, Saarland University) – CGH Data Analysis
Prof. Zazzi (Virology, University of Siena) – Bioinformatics for HIV
Prof. Zeuzem (Gastroenterology, University of Frankfurt) – HCV

We are cooperating within the following scientific consortia:

Arevir Consortium – HIV Resistance Patterns
BioSapiens Network of Excellence – Structure-function Relationships in Proteins/Bioinformatics for Infectious Diseases
DFG Clinical Research Group on Hepatitis C – HCV
Eu*Resist* – European STREP on HIV resistance analysis
National Genome Research Network NFGN: SMP Bioinformatics – Expression Data Analysis
National Genome Research Network NFGN: Genomic Network on Environmental Diseases – Analysis of the Structure and Function of Medically Relevant Proteins

We are cooperating with the following company:

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

Software policy

It is the acclaimed intention of the group to develop bioinformatics software that is useful to a wide user community. Thus the group also commits to realizing channels of dissemination and evaluation of the software, be it commercial or non-commercial. For instance, our protein structure prediction software (ARBY Server) and the analysis software of HIV genotypes for drug resistance (Geno2Pheno Server) has been made accessible non-commercially via an Internet server offer. Other software 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 have been made available via our web server (see `http://www.mpi-sb.mpg.de/units/ag3/software.html`). These include

BIQ Analyzer is a free software tool for easy visualization and quality control of DNA methylation data. With more than a thousand downloads so far, BiQ Analyzer has become a standard tool for processing DNA methylation data from bisulfite sequencing. BiQ Analyzer has been selected by ABI to be part of the Applied Biosystems Software Community Program and published in an applications note to Bioinformatics [3].

DomainNetworkBuilder/NetworkAnalyzer are two Java plugins for Cytoscape, a free open-source software platform for visualization and analysis of biomolecular networks. DomainNetworkBuilder decomposes protein networks into domain-domain interactions and generates a new network of interacting domains and is described in a short paper in ECCB 2005 [1]. NetworkAnalyzer computes parameters describing the network topology and displays their distributions in diagrams.

GOTax is a free comparative genomics platform that integrates protein annotation with protein family classification and taxonomy. User-defined sets of proteins, protein families, annotation terms or taxonomic groups can be selected and compared, allowing for the analysis of distribution of biological processes and molecular activities over

different taxonomic groups. Additionally, a functional similarity measure is available for establishing functional relationships between proteins and protein families. GOtax has been published in Genome Biology [8].

IRECS is a software package for side-chain placement, which is especially tailored for the needs of molecular docking. In contrast to other side chain placement tools, our tool is able to predict an ensemble of the most probable conformations for each side chain of a protein depending on its flexibility [5].

NOXclass is a classifier identifying protein-protein interaction types (biological obligate, biological non-obligate and crystal packing) implemented using a support vector machine (SVM) algorithm. NOXclass has been published in BMC Bioinformatics [9].

Recco analyzes alignments of sequences that evolved subject to recombination and mutation. The analysis provides evidence as to whether a dataset contains recombination, which sequence is a recombinant and where the recombination breakpoints are. The analysis is based on explaining one sequence with all other sequences in the alignment using mutation and recombination. A parametric analysis of the parameter alpha, which weights recombination cost against mutation cost, yields additional information as to which sequence might be recombinant. Recco is freely available for download and has been published in Bioinformatics [7].

topGO (topology-based Gene Ontology scoring) is a software package for calculating the significance of biological terms from gene expression data. It implements various standard and advanced new algorithms for determining the relevance of Gene Ontology groups from microarrays. A specific feature of the advanced algorithms is the exploitation of the hierarchical graph structure of the GO annotation for coping with the large number of GO groups. Often, related biological terms are scored with a similar statistical significance. Dependencies between GO terms can be de-correlated by accounting for the neighborhood of a GO node when calculating its significance. The new algorithms better detect significant GO terms from gene expression data. topGO is freely available for download and has been published in Bioinformatics [2].

Among the new packages, the most requested ones are BIQ Analyzer and Network Analyzer, with downloads in the hundreds and Recco with downloads in the dozens. Among the older packages geno2pheno continues to be popular with over 40000 hits since December 2000, and ROCR is continuously distributed via the CRAN repository.

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 group. We have allocated a full scientist (Joachim Büch) 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 group 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 (14.9.1).

References

- [1] M. Albrecht, C. Huthmacher, S. C. Tosatto, and T. Lengauer. Decomposing protein networks into domain-domain interactions. *Bioinformatics*, 21(Suppl. 2):ii220–ii221, 2005.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] C. Hartmann, I. Antes, and T. Lengauer. IRECS: A new algorithm of the selection of most probable ensembles of side-chain conformations in protein models. *Protein Science*, 15, 2007.
- [6] T. Lengauer and T. Sing. Bioinformatics-assisted anti-HIV therapy. *Nature Reviews Microbiology*, 4:790–797, October 2006.
- [7] J. Maydt and T. Lengauer. Recco: recombination analysis using cost optimization. *Bioinformatics*, 22(9):1064–1071, February 2006.
- [8] 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, March 2007.
- [9] H. Zhu, F. S. Domingues, I. Sommer, and T. Lengauer. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7:1–15, June 2006.

6 The Computer Graphics Group (D4)

The computer graphics group has been established in 1999 and currently consists of about 40 researchers. There has again been a considerable change in personnel over the past two years, and seven more researchers have received offers for faculty positions during the reporting period. This brings the total number of researchers from the former Erlangen/ now Saabrücken group who have received offers for faculty positions since 1999 to a total of 22 persons [1]. Four more researchers accepted positions for postdoc positions from leading institutions abroad.

Despite this loss of expertise, the current group has performed extremely well during the reporting period, and members of the group have again co-authored a multitude of papers in the premier journals (ACM TOG, IEEE TVCG, IJCV, GMOD, CAGD, CGF, Visual Computer, IJSM, etc.) and conference venues (ACM SIGGRAPH, Eurographics, EGSR, SGP, SMI, ACM SPM, IEEE Visualization, NPAR, Pacific Graphics, ICCV, ECCV, CVPR, VMV, DAGM, etc.) in the field.

Vision and Research Directions

During the last two decades computer graphics has established itself as a core discipline within computer science and information technology. Computers are more and more used to model, simulate and render parts of the real or an imaginary world, and due to the importance of visual information for humans, computer graphics is at the very core of the technologies enabling the modern information society. 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, if possible interactive, high quality visual display of the results. 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

[1] T. Ertl (C4, Stuttgart), P. Slusallek (C4, Saarbrücken), G. Greiner (C4, Erlangen), L. Kobbelt (C4, Aachen), R. Westermann (C4, Munich), J. Haber (C4, Dresden, declined), A. Kolb (C4, Siegen), S. Gumhold (W3, Dresden), M. Stamminger (C3, Erlangen), H. Theisel (W2, Bielefeld), V. Blanz (W2, Siegen), U. Schwanecke (C3, FH Wiesbaden), W. Heidrich (Assoc. Prof., Univ. British Columbia, Canada), J. Kautz (Lecturer, University College, London), P. Bekaert (C3, Univ. Limburg, Belgium), C. Soler (Tenured Researcher, INRIA, France), S. Ghali (Ass. Prof., Univ. Alberta, Canada), J. Lang (Ass. Prof., Univ. Ottawa, Canada), S.-W. Choi (Research Fellow, KIAS, Korea), I. Ivrissimtzis (Lecturer, Durham University, UK), M. Neff (Ass. Prof, UC Davis, USA), V. Havran (Lecturer, TU Prague, Czech Republic)

simulation to compute an image, we like to take the more integrated view of *3D Image Analysis and Synthesis* for our research, and consider the whole pipeline from data acquisition over 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) side. According to this point of view, one of the key research challenges then is the development of good models and modeling tools to efficiently handle the huge amount of data during the acquisition process and to facilitate further processing and rendering.

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 Topics and Structure of the Group

As mentioned above we consider the whole pipeline from data acquisition over 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 nine research areas, each having its own small group of coordinators:

- Digital Geometry Processing (A. Belyaev)
- Freeform Surfaces and Visualization (C. Rössl)
- Vector Field Visualization and Applications of Vector Field Techniques (H. Theisel)
- Modeling and Animation of Faces (V. Blanz)
- Markerless Motion Capture (B. Rosenhahn)
- Multiview Video Processing (C. Theobalt)
- General Appearance Acquisition and Computational Photography (H. Lensch)
- Advanced Global Illumination and Realtime Realistic Image Synthesis (J. Günther and K. Myszkowski)
- High Dynamic Range Imaging and Perception Issues in Graphics (R. Mantiuk and K. Myszkowski).

The coordinators coordinate the work in their areas and together with Hans-Peter Seidel form the AG4 steering committee. The steering committee meets on a weekly basis 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.

Some Achievements

We have been pursuing first-class research on a broad range of topics, and members of the group have actively published in the top conferences and journals (45 journal articles, 126 conference articles) (see Section 15.18 for details). We have actively participated in program committees and have given numerous invited talks and tutorial presentations at major national and international events (see Section 15.14 for details). Our software has been successfully integrated and validated in a variety of projects (see Sections 15.13 and 15.17), and researchers from the group have spread out to other institutions.

In the following we briefly highlight some of our achievements in each of the ten research areas:

Digital Geometry Processing

We developed new algorithms for shape reconstruction based on multilevel partition of unity implicit radial basis functions (GMOD'05, GMOD'06), on template based methods (SPBG'06), and on the use of ensembles and statistical learning (CAD'07, TVC'07), and for denoising (ACM SPM'07). We contributed to the theory of discrete surface parameterization (SGP'05, IJSM'05, IJSM'06), of generalized barycentric coordinates (SGP'06), and on curvature tensor estimation (CAGD'07), and we worked on shape annotation and view selection (SMI'06).

Free-Form Surfaces and Visualization

During the reporting period we mostly focused on shape editing (IEEE VR'07), shape deformation (EG'05) and the use splines and higher order primitives for efficient visualization. This involves the use of higher order primitives (IEEE Vis'05, IEEE RT'06), and a thorough study of the approximation and interpolation properties of certain trivariate splines (CAGD'05).

Vector Field Visualization and Applications of Vector Field Techniques

Topological methods for vector field visualization can be used to segment the vector field into regions of different flow behavior. We continued our previous work on multifield visualization (IEEE Vis'06), extraction of surfaces in time dependent vector fields (IEEE Vis'05) and path line oriented flow topology (EuroVis'06), and successfully addressed segmentation of DT-MRI anisotropy isosurfaces (EuroVis'07).

Another rather surprising application of vector field techniques was an algorithm for vector field based shape deformations. The resulting deformation is volume-preserving, feature and smoothness preserving, and guaranteed to be free of (local and global) self-intersections (ACM SIGGRAPH'06).

Modeling and Animation of Faces and Hands

The goal of learning-based modeling is to replace much of the manual design by automated procedures for measurement and data analysis. During the reporting period we have applied

our framework to the creation of face models from vague mental images (EG'06), and to learning-based facial rearticulation using streams of 3D scans (PG'06). In cooperation with researchers at MPI for Biological Cybernetics we have also applied machine learning for the processing of raw 3D scan data (EG'05, ICML'05).

Animated characters that move and gesticulate appropriately with spoken text are useful in a wide range of applications. We have developed a system that is capable of producing gesture animation for given input text in the style of a particular performer (extracted from existing video footage) (ACM TOG'07).

Markerless Motion Capture

The subgroup on markerless motion capture deals with open questions regarding modeling, tracking, understanding and analyzing human motions from video data (DAGM'05, KI'06). Results during the reporting period include an algorithm for pose estimation of free-form contours (ICCV'05), an algorithm for 3D pose tracking by exploiting contour and flow constraints (ECCV'06), an algorithm for joint image segmentation and pose estimation by exploiting 3D shape knowledge (DAGM'05), and a system for articulated tracking incorporating a clothing model (DAGM'06, MVA'07)

Multiview Video Processing and Vision-based Computer Graphics

The goal of multiview video processing is to generate realistic renderings of image- or video-captured dynamic scenes from arbitrary virtual cameras. During the reporting period we have extended our original model-based free-viewpoint video approach by a concept called dynamic reflectometry which enables us to recover time-varying surface reflectance properties from eight input video streams captured under calibrated lighting. This enables us to render 3D videos in real-time from arbitrary viewpoints and under arbitrary illumination (IEEE TVCG'07). Another important result is our algorithm for markerless deformable mesh tracking for human shape and motion capture (CVPR'07).

General Appearance Acquisition and Computational Photography

The goal of this group is to develop techniques and algorithms to capture realistic models of real-world objects and phenomena. The acquired models then allow to synthesize novel views of the digitized objects, or to change the incident illumination virtually. Specific results include an algorithm for model-based reflectance for human faces (IEEE TVCG'05), implicit reflectance with Bayesian relighting (EGSR'05), and our dual photography techniques for capturing reflectance fields which can be correctly relit even with spatially varying light patterns and nearby light sources (ACM SIGGRAPH'05, EGSR'06). Other important results are our algorithm for mesostructure reconstruction from specularity (CVPR'06), and our work on automatic acquisition of multiperspective street panoramas (EGSR'06).

Advanced Global Illumination

Ray tracing and global illumination techniques significantly improve the realism of synthetic images. Both the hardware and algorithms at hand are now mature enough to make realistic

rendering of animated environments feasible even at interactive speeds. Specific results during the reporting period include a new algorithm for computing high quality indirect illumination based on photon density estimation (EG'05), faster isosurface ray tracing using implicit kd-trees (IEEE TVCG'05), and interactive ray tracing of dynamic scenes (EG'06, TVC'06, IEEE IRT'06).

New global illumination and rendering solutions that exploit temporal coherence in lighting distribution for subsequent frames can substantially improve both the computation performance and overall animation quality (EGSR'05, EG'05). Finally, we developed a new physically-based simulation of twilight phenomena (ACM TOG'05).

High Dynamic Range Imaging and Perception Issues in Graphics

Due to rapid technological progress in high dynamic range (HDR) video capture and display, the efficient storage and transmission of such data is crucial for the completeness of any HDR imaging pipeline. We have continued our successful work on HDR imaging and developed algorithms for backward compatible high dynamic range MPEG video compression (ACM SIGGRAPH'06), a perceptual framework for contrast processing of high dynamic range images (ACM TAP'06), and improved tone mapping operators for HDR images and video (EG'05, EG'06). We also conducted an analysis of reproducing real-world appearance on displays of varying dynamic range (EG'06).

Prizes and Awards

Several group members have received awards for their work during the reporting period. Hendrik Lensch received the Eurographics Young Researcher Award 2005, Rafal Mantiuk received the Heinz Billing Prize 2006, and Bodo Rosenhahn received the DAGM Main Prize 2005. Christian Theobalt received the Eduard Martin Award 2006 (PhD award by Saarland University), Michael Goesele received the Eduard Martin Award 2005, and Martin Fuchs (VDI Prize 2005) and Jürgen Gall (SMI-DAGM Graduation Award) received awards for their master theses. Ingo Wald received the SaarLB Wissenschaftspreis 2005. Finally, Grzegorz Krawczyk, Waqar Saleem, and Bodo Rosenhahn all received best paper awards at different conferences.

Cooperations

In addition to the collaborations inside the institute (surface reconstruction, free viewpoint video) and with the university (realtime 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 substantial number of formal cooperations with other institutions on both a national and international level. On a European level, these includes the EU projects RealReflect, Aim@Shape (NoE), and 3DTV (NoE). We also collaborate directly with Daimler Chrysler (simulation of display appearance in the car cockpit)

and Dolby Imaging Labs (High Dynamic Range Video). Details on those cooperations can be found in Section 15.17.

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 Informatics 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 Informatics and ultimately as a professor at a German university.

The center is directed jointly by Hans-Peter Seidel (MPII) and Bernd Girod (Stanford) and currently encompasses twelve independent research groups (six at Stanford and six at MPII).

Group Development

The composition of the group has changed considerably over the past two years.

Bodo Rosenhahn joined us as leader of a junior research group within the Max Planck Center for Visual Computing and Communication (MPC-VCC) in 2005. Hendrik Lensch returned from his postdoc at Stanford in 2006 and also became junior research group leader within MPC-VCC. Hendrik has also been awarded an Emmy-Noether-Fellowship by the German Research Council (DFG). Rafal Mantiuk successfully defended his PhD in 2006 and continues as a postdoc.

Stefan Gumhold (W3-Professorship, TU Dresden), Volker Blanz (W2-Professorship Univ. Siegen), Holger Theisel (W2-Professorship Univ. Bielefeld), Jan Kautz (Lecturer, University College London), Ioannis Ivrissimtzis (Lecturer Durham University, UK), Michael Neff (Tenure Track Assistant Professor UC Davis, USA), and Vlastimil Havran (Lecturer TU Prague, Czech Republic) all accepted offers for faculty positions and left the group.

Michael Goesele (Univ. Washington), Christian Rössl (INRIA Sophia Antipolis), Christian Theobalt (Stanford), and Ingo Wald (Univ. Utah) accepted offers for postdoc positions from leading institutions abroad. Jörg Haber joined a special high tech task force with the Bavarian police, and Irene Albrecht, Kirill Dmitriev, Zachi Karni, and Hitoshi Yamauchi accepted senior positions in industry.

Our current PhD students are Naveed Ahmed, Boris Ajdin, Thomas Annen, Tunc Aydin, Robert Bargmann, Tongbo Chen, Edilson De Aguiar, Zhao Dong, Christian Fuchs, Martin

Fuchs, Jürgen Gall, Johannes Günther, Nils Hasler, Robert Herzog, Matthias Hulling, Grzegorz Krawczyk, Torsten Langer, Waqar Saleem, Natascha Sauber, Oliver Schall, Kristina Scherbaum, Thomas Schultz, Kuangyu Shi, Kaleigh Smith, Wenhao Song, Carsten Stoll, Martin Sunkel, Wolfram von Funck, Akiko Yoshida, Rhaleb Zayer, and Gernot Ziegler.

7 The Databases and Information Systems Group (D5)

Overview

D5 has been established in October 2003. It is headed by Gerhard Weikum, and currently consists of 14 doctoral students and 7 post-doctoral researchers. The group's research falls into four major areas each of which is coordinated by a research associate or senior postdoc:

1. semantic search and semistructured data management (coordinated by Ralf Schenkel),
2. peer-to-peer search and information management (coordinated by Christos Tryfonopoulos),
3. query processing and optimization (coordinated by Thomas Neumann), and
4. Web and text mining (coordinated by Srikanta Bedathur).

Vision and Research Directions

The group's research activities are inspired by and strive for the following visions and long-term goals.

Intelligent Organization and Search of Information

The age of information explosion poses tremendous challenges regarding the intelligent organization of data and the effective search of relevant information in business and industry (e.g., market analyses, logistic chains), society (e.g., health care), and virtually all sciences that are more and more data-driven (e.g., gene expression data analyses and other areas of bioinformatics). The problems arise in intranets of large organizations, in federations of digital libraries and other information sources, and in the most humongous and amorphous of all data collections, the World Wide Web and its underlying numerous databases that reside behind portal pages. The Web bears the potential of being the world's largest encyclopedia and knowledge base, that would be of great value to all kinds of "knowledge workers" from students to Nobel laureates, but we are very far from being able to exploit this potential.

Search-engine technologies provide support for organizing and querying information; for simple mass-user queries that aim to find popular Web pages on pop stars, soccer clubs, or the latest Hollywood movies, existing engines like Google are probably the best solution. But for advanced information demands search engines all too often require excessive manual preprocessing, such as manually classifying documents into a taxonomy for a good Web

portal, or manual postprocessing such as browsing through large result lists with too many irrelevant items or surfing in the vicinity of promising but not truly satisfactory approximate matches. The following are example queries where current Web and intranet/enterprise search engines fall short:

Q1: Which professors from Saarbruecken in Germany teach information retrieval and participate in EU projects?

Q2: Which drama has a scene in which a woman makes a prophecy to a Scottish nobleman that he will become king?

For Q1 no single Web site is a good match; rather one has to look at several pages together within some bounded *structural context*: the homepage of a professor with his address, a page with course information linked to by the homepage, and a research project page that is a few hyperlinks away from the homepage. Q2 cannot be easily answered because a good match does not necessarily contain the keywords "woman", "prophecy", "nobleman", etc., but may rather say something like "Third witch: All hail, Macbeth, thou shalt be king hereafter!" and the same document may contain the text "All hail, Macbeth! hail to thee, thane of Glamis!". So this query requires some *background knowledge* to recognize that a witch is usually female in the literature, "shalt be" refers to a prophecy, and "thane" is a title for a Scottish nobleman.

Answering such queries with high precision, despite high diversity and noise in the underlying data, calls for a new kind of *"semantic search" engine* that integrates concepts and techniques from database (DB) and information-retrieval (IR) systems. XML IR falls into this framework, especially when combined with background knowledge in the form of ontologies and when the underlying data has semantically expressive tags. Graph IR is a generalization towards entity-relation graphs that capture highly non-schematic RDF or cross-linked XML data or relational facts that are automatically extracted from text and Web sources. Such advanced DB&IR approaches pose major challenges regarding the quality of search results (in terms of IR measures like precision and recall) and the efficiency and scalability of indexing and query processing.

Self-organizing Distributed Information Systems

Even when we focus on scientific information alone and leave out business and entertainment data, new information is produced and compiled world-wide in a highly distributed manner, in federations of digital libraries or e-science repositories and, of the course, on the Web with millions of scholars and students. Thus, it is natural to pursue a completely decentralized peer-to-peer (P2P) architecture for managing this information explosion, intelligently organizing the information, and efficiently searching it.

The P2P approach bears the potential of overcoming the shortcomings of today's Web search engine technology and successfully tackling the kinds of killer queries mentioned before. In our architecture every peer, e.g., the home PC of a scientist or student, has a full-fledged search engine that indexes a small portion of the Web, according to the interest profile of the user. In addition to conventional on-demand search, this architecture could also be better suited for advanced services like multimedia search or publish-subscribe services.

For example, users search for photos or video clips in combination with user-provided tags and textual or speech comments, or a user registers her continuous information demand and will be automatically notified whenever new data is posted that satisfies her needs. In both settings, various facets of social networks among users could be exploited for better quality and efficiency. Finally, another challenging scenario would be to build and maintain an Internet archive with the version history of all Web sites for long-term preservation and "time-travel" search, using a P2P system approach for scalability and availability.

A P2P architecture has four major advantages over a centralized server farm:

1. As the data volume and the query load per peer are much lighter, the peer's search engine can employ much more advanced techniques for concept-based rather than keyword-based search, leveraging background knowledge in the form of thesauri and ontologies and powerful mathematical and linguistic techniques such as spectral analysis and named entity recognition.

2. Peers can collaborate for finding better answers to difficult queries: if one peer does not have a good result locally it can contact a small number of judiciously chosen peers who are considered "knowledgeable" on the query topic. This approach should often be able to exploit the small-world phenomenon on the Web: knowledgable peers are only a short distance away.

3. A P2P system can gather and analyze bookmarks, query histories, user click streams, and other data about user and community behavior; the implicit and explicit assessments and recommendations derived from this input can be leveraged for better search results. In contrast to a central server, the P2P approach provides each user with direct, fine-grained control over which aspects of her behavior may be collected and forwarded to other peers.

4. A politically important issue is that a P2P search engine is less susceptible to manipulation, censorship, and the bias induced by purchased advertisements.

Thus, a P2P approach to information search could pave the way towards a *"social search"* super-engine that leverages the collaborative "wisdom of crowds". Challenges that we face towards this vision are scalability and the desire to make the P2P network self-organizing and resilient to high dynamics (failures and churn) and possible manipulation (cheating, spam, etc.).

Research Areas

Each of the two visions – intelligent search and self-organizing P2P information management – is pursued in one of the four research areas that the group is working on: *semantic search and semistructured data management*, coordinated by Ralf Schenkel, and *peer-to-peer search and information management*, coordinated by Christos Tryfonopoulos.

The envisioned "semantic search" should not only address the querying itself, but also aim at *intelligent organization of information* because data with more explicit structure, semantic annotations, and clean data items becomes easier to search with high precision. But the Web and also federations of data sources are highly diverse in terms of structure, annotations,

and data quality (e.g., authority, resolution, completeness, freshness, etc.). Likewise, "social search" can benefit from a deeper understanding of the underlying user-behavior patterns and the interactions and other relations among users. This requires the analysis of link graphs and similar interconnection structures, by using spectral analysis techniques and other graph mining methods. For these reasons, we also pursue research in the area of *Web and text mining*, combining techniques from data mining, natural language processing, and statistical machine learning. This area is coordinated by Srikanta Bedathur.

Last but not least, all these goals can only be achieved, with practically viable methods, when paying attention to the efficiency and scalability of the underlying algorithms. This holds most notably for indexing and query processing, especially for top-k queries that are needed for ranked retrieval and other kinds of aggregation and "iceberg" queries. Therefore, we also pursue research in the area of efficient *query processing and optimization*, which is coordinated by Thomas Neumann.

Achievements

The group has established itself as one of the world-wide leading research groups on DB&IR integration, with first-rate work in both scientific communities and considerable achievements on the fusion of both technologies, most notably on semantic XML IR and graph IR. The work on P2P search has received wide recognition as well. Finally, the expansion of the group's original expertise into the areas of Web and text mining has been successful. In all these areas, many of our scientific results have been integrated into software prototypes (TopX, Minerva, YAGO/NAGA, TTIX). More specifically, highlight results include:

- algorithms for efficient top-k query processing for text, XML, and distributed index lists (partly in collaboration with D1, and integrated into TopX),
- new forms of peer synopses for correlation-aware query routing in P2P systems (integrated into Minerva),
- a general algorithm for decentralized, scalable link analysis with guaranteed convergence and strong system properties,
- a new approach to knowledge harvesting from Web sources and their organization and effective search in knowledge graphs (as part of the YAGO ontology and the corresponding search engine NAGA),
- the indexing and query processing methods for "time-travel" search on Web-history and other text-version collections (integrated in the TTIX software).

The group puts major efforts into prototyping software systems, and pursues the philosophy of open source software dissemination. Currently, the focused crawler BINGO!, the XML IR system TopX, and the P2P search platform Minerva can be freely downloaded from the group's Web site. BINGO! has also been integrated with the Minerva platform, and this combination is used by project partners in the EU projects DELIS and DELOS. TopX has served as the reference engine for topic development of the 2006 INEX benchmark, which involved more than 50 international research groups on XML IR. Minerva and TopX have been

shown in the (refereed and selective) demo programs of several major conferences (VLDB 2005, Middleware 2005, CIDR 2007, SIGMOD 2007) and received very positive feedback. The software for constructing the YAGO ontology and its corresponding graph-based search engine NAGA are planned to be released soon, and the time-travel search engine TTIX for text and Web archives is being prepared for open-source availability as well. All our software systems are also intensively used for new projects within the group itself; for example, several doctoral students have used BINGO! for collecting and preparing data in their experimental studies, and TopX is being used for new work on proximity search, relevance feedback, and semantically enhanced XML querying.

The group emphasizes the experimental evaluation of newly developed models, algorithms, and systems. Substantial efforts have been made to build up appropriate datasets (e.g., Web crawls, Wikipedia in XML format with additional annotations, ontologies mined from WordNet and Wikipedia, Web version data from the European Archive, etc.) and the necessary infrastructure for experimentation. The group has been participating in various tracks of the TREC benchmark conference and the INEX benchmark for XML information retrieval.

Several group members have won awards: Fabian Suchanek received the Günter-Hotz Medal of the Saarland University for the best Master theses; Martin Theobald received the Dissertation Award of the German Computer Society's Chapter on Database and Information Systems (GI DBIS) and has been selected for the 2007 Otto-Hahn Medal of the Max-Planck Society and an Honorable Mention Award (i.e., on of the top three) for the 2007 ACM SIGMOD Dissertation Award; Gerhard Weikum has become an ACM Fellow in 2005.

The group has been very successful also in terms of first-rate publications: 8 full papers in the VLDB, EDBT, ICDE, SIGIR, CIKM conferences of the year 2006, which are among the premier conferences in the database and information retrieval fields and have acceptance ratios of 20% or lower, and numerous papers in other highly selective conferences and workshops (such as PKDD, IPTPS, ECIR, WISE, and WebDB, all of which have acceptance rations of 25% or lower). A more specific quality indicator is the publications at VLDB, one of the two top-rated conferences in the database systems area, and SIGIR, the premier conference in the information retrieval area: in the three years 2004 through 2006 we had 8 full papers and 2 demo papers in VLDB, and in the three years 2005 through 2007 we had 5 full papers in SIGIR. Our paper at CIKM 2006 has won the Best Interdisciplinary Paper Award (which is the conference's best paper award, as the conference aims at bridging the DB and IR communities), and a paper at VLDB 2006 has been selected for a best-of-VLDB'06 special issue of the VLDB Journal.

The group has built up various relations with industry, both IT companies, especially in the search engine market, and application businesses. Connections have been made with Google, Yahoo!, and Ask through several mutual visits. The group has two joint publications with researchers of the Yahoo! Research Lab in Barcelona, and two doctoral students will spend three-month internships with the Google Lab in Zurich. Closer collaborations have been established with the IBM Research Lab in Haifa, Israel, with the European Archive in Amsterdam, the European sibling of the Internet Archive whose mission is the digital preservation of Internet contents, and with The European Library (TEL) that provides a portal to Europe's National Libraries.

Collaborations and Networking

Within the institute, D5 is primarily cooperating with D1 (especially Holger Bast's group) on efficient algorithms for information retrieval. The two groups jointly participate in the EU-funded Integrated Project DELIS on Dynamically Evolving Large-Scale Information Systems, and they have co-authored several papers including a VLDB 2006 and a SIGIR 2007 paper. There are various smaller-scale local collaborations with bioinformatics (Lenhof at Saarland University) on rule mining for gene expression time-series, material sciences (Mücklich at Saarland University) on automatic classification of microstructure images, computational linguistics (Uszkoreit at DFKI, Pinkal at Saarland University) on richer representations of natural-language text, computational logics (Weidenbach at RG1, Melis at DFKI) on ontologies, machine learning (Scheffer at RG2, Zeller at Saarland University) on mining Web-usage logs (queries, clicks) and software development histories, security and cryptography (Backes at Saarland University) on anonymity and censorship resistance in P2P networsk, and with the Dagstuhl Seminar Center (Wilhelm) on using advanced link analysis techniques for impact assessment.

At the national level the group collaborates with the information retrieval group of the University of Duisburg (Fuhr) on the DFG-funded project CLASSIX (classification and search of information in XML), with the Heinz-Nixdorf Institute in Paderborn on P2P Web search in the context of the EU project DELIS, and with the Technical University of Berlin (Feldmann) on analyzing search-engine traffic logs.

Internationally, the group participates in the EU Integrated Projects DELIS and AEOLUS, with intensive collaboration with the University of Patras, Greece; the EU Network of Excellence DELOS on future digital libraries, including a joint task with The European Library (TEL) in The Hague; and the newly started EU Specific Target Project SAPIR, with collaboration with IBM Research in Haifa, Telenor in Oslo, CNR in Pisa, and various other partners. In addition the group informally collaborates with the Athens University of Economics and Business (Vazirgiannis), the University of Athens (Koubarakis), Microsoft Research in Redmond (Chaudhuri, Lomet), and Yahoo! Research in Barcelona (Castillo, Donato), all of which produced joint publications; and it maintains close contacts to various other universities and companies (e.g., IBM, Google, SAP, Telenor, Unilever, Elsevier, Juris).

Group Development

After the initial build-up of the group in the years 2004 and 2005, the group is now in steady state and has a healthy balance of graduate students and more experienced research associates and postdocs.

Major strengths are the fact that the group has competence 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. As for institute-internal collaboration, strong ties have been developed with D1 on algorithmic issues of information retrieval, and a similarly strong connection is aimed for with the newly established RG2 on machine learning. Major opportunities on the near-term horizon are technology transfers via open-source software like TopX and Minerva, and the

collaboration with Internet and digital-library organizations like the European Archive and The European Library. The theme of knowledge harvesting and semantic search (as started in the work on YAGO and NAGA) is expected to become a major focus with challenging long-term perspectives.

A potential threat is the group's diversity of research topics and the fluctuation of researchers. There are continuous efforts (weekly meetings, reading groups, etc.) to keep the group reasonably coherent, and the director and the group 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, it is considered healthy for graduates and beneficial for their careers that they move to other research organizations. Thus, a certain extent of staff dynamics seems unavoidable.

8 The Automation of Logic Group (RG1)

The Automation of Logic Group has been established in September 2005 and is headed by Christoph Weidenbach, who rejoined the Max Planck institute after working as an IT manager at General Motors Europe for 6 years. The group has its roots in the Programming Logics Group. Actually, all members are former members of Harald Ganzinger's Programming Logics group, except for Matthias Horbach and Manuel Lamotte who joined the group in 2006 and 2007, respectively. The group is following the tradition of the Programming Logics group with an even stronger focus on showing practical relevance of the developed results. Almost all of our established theories, procedures and reasoning technology is implemented in software prototypes and evaluated in practice, either by us or cooperating groups.

Vision and Research Directions

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. In particular, we are interested in combination procedures motivated by hierarchic or modular theory reasoning, decision procedures in the superposition context, advances in first-order model checking, applications to verification problems and the further development of our software.

Hierarchic and Modular Reasoning

Complexity in reasoning often arises from combinations of different theories. Typical relationships between theories are hierarchical in the sense that one theory is a sub-theory of the other, or modular for combinations of theories, possibly sharing a common subtheory.

A promising approach to automatize reasoning in such theory combinations is to reflect the theories relations by the reasoning mechanisms. We successfully developed superposition calculi for modular theory combinations, and reduction and interpolant generation methods for hierarchical and in particular local theory extensions. We investigated model theoretic properties of systems expressed by Cartesian axioms, and started to look into reductions from inductive theorem proving to first-order saturation.

Decision Procedures

Decision procedures are at the core of the automation of logic. They come as refinements of existing procedures or calculi as well as special purpose algorithms in order to effectively decide certain logic fragments.

For a number of relevant theories the superposition calculus can actually be turned into an effective and efficient decision procedure that can compete with special purpose systems. We presented finitely saturating classes for the bit vector theory and developed a refinement

of the superposition calculus deciding formulas over finite domains. A consequence of the latter is the superposition decidability of the Bernays-Schönfinkel class.

We developed decision procedures for several types of unification problems for distributive lattices.

First-Order Model Checking

Model checking is currently one of the most relevant technologies for the automatic analysis and verification of systems. For any model checking approach, efficient reasoning mechanisms capturing the typically huge state space are key for any practical applicability.

We investigated hybrid systems with large discrete state spaces where the continuous part is represented by first-order formulae and the discrete part is modeled in a symbolic way. The symbolic representation supports the aggregation of states and hence enables efficient reasoning technology for large discrete state spaces.

Applications

The application of our results is an inherent part of our research. It both gives us feedback on the impact of our methods and identifies new lines of research.

Our results on hierarchical and modular reasoning enabled us to to successfully address a series of case studies in which (i) safety properties of systems with a parametric number of components could be verified, and (ii) constraints between certain parameters of such systems could be established ensuring that safety conditions are fulfilled. The systems we considered consist of an undetermined (parametric) number of trains on a linear track controlled by a radio control center. 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. Moreover, in a parametric setting, we entail constraints between minimum and maximum speeds, and the time intervals for communication which guarantee safety of the system.

The algorithm for hierarchical computation of interpolants was applied in the context of static software analysis. Here leading tools are based on the abstraction refinement principle and need to efficiently compute interpolants for extensions of linear arithmetic with free function symbols, perfectly fitting our results. Our implementation was successfully integrated into the predicate discovery procedure of the software verification tools Blast and ARMC.

The configuration, maintenance and debugging of IT-infrastructures is a non-trivial task. Our approach is to analyze infrastructures with respect to defects, e.g., bugs in the configuration that prevent a client from obtaining a service. We developed an integrated first-order model that starts at Ethernet level and can handle all IT-infrastructure functionality up to higher level protocols such as DHCP. The model is finitely saturated by SPASS and we instantiated it successfully for the overall LAN infrastructure of both Max Planck institutes in Saarbrücken.

We added labels to first-order clauses to simultaneously apply superposition to several proof obligations inside one clause set. Motivated by software verification, we have created a prototype implementation of labelled clauses that supports multiple conjectures, and we could provide convincing experiments for the benefits.

Software

We currently develop three main software tools: SPASS, SPASS+T and Waldmeister. SPASS is one of the leading saturation procedures for full first-order logic with equality. Recently, we extended it by translation mechanisms for non-classical logics and enhanced application interfaces.

The prover SPASS+T is currently the only available system that supports full first-order logic with equality plus (linear) arithmetic. The potential of the superposition based reduction technology is used for the simplification of formulae including (linear) arithmetic expressions that are eventually passed to a (linear) arithmetic solver.

The unfailing completion procedure Waldmeister, worldwide dominates this area now for over 10 years. In particular, it provides support for the computation of representation theorems in weak algebraic structures, like groups or semi-groups.

Projects and Collaborations

Viorica-Sofronie Stokkermans and Uwe Waldmann took over part of Harald Ganzinger's projects in the AVACS research center and became principal investigators of the consortium in 2004. Together with Christoph Weidenbach they are principal investigators for the second phase proposal of the transregional collaborative research center AVACS (Automatic Verification and Analysis of Complex Systems) linking the sites Oldenburg, Freiburg and Saarbrücken and funded by the German Research Foundation (DFG) starting in 2008. Most of our research is devoted to this project.

Until mid of 2006, we were involved in the Verisoft research project funded by the German Federal Ministry of Education and Research. The main goal of the project is the pervasive formal verification of computer systems.

Our software systems SPASS, SPASS+T, and Waldmeister are used by more than hundred users/groups around the world. The collaboration with these users/groups varies from loose contact to the execution of common research plans. Currently, we have intensive contacts to the static verification group of Mooly Sagiv at Tel Aviv, Israel, the CASL group of Bernd Krieg-Brückner at Bremen, Germany, and the higher-order logic theorem proving (Isabelle) group of Larry Paulson at Cambridge, UK.

Inside the MPI we have started a master thesis project together with Kurt Mehlhorn's group on the integration of linear arithmetic optimization procedures into SPASS and are planning a project with Gerhard Weikum's group on efficient reasoning mechanisms for the semantic web.

Group Development

Basically, the group is intended to stay at its current size. Depending on the success of our running project applications, it may increase by one or two positions in 2008.

Out of the INRIA–MPI postdoc agreement, Duc Khanh-Tran from INRIA Nancy will join the group in October 2007 for a one year stay.

9 The Machine Learning Group (RG2)

The Machine Learning Research Group has been established in January 2007; it is headed by Tobias Scheffer. The group currently consists of eight doctoral students and four Master's students. The group's research aims at understanding how algorithms can be constructed that effectively learn from data to perform better at their tasks. A part of the group's research is dedicated to applications of machine learning in the areas of natural language processing and information retrieval. In the next years, the group plans to also pursue cooperations that address problems in bioinformatics and software engineering.

Vision and Research Direction

The group's application-oriented line of work is 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 personalized spam filtering and personalized relevance ranking falls into this category.

The research group's technology-oriented research focuses on abstract, more general problem setting 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 structural learning falls into this category.

Learning from Differing Training and Test Distributions

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. 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 geographic regions where vegetation and lighting conditions result in a distinct distribution of input patterns.

The Machine Learning Group has contributed several results to this area. A logistic regression classifier for differing training and test distributions is among these results. The training algorithm considers both training and test data, and comes to a classification hypothesis that performs well with respect to the test data. One can show that this classifier is the *Maximum A Posteriori* solution with respect to the test distribution; in other words, the classifier is an optimal solution in a rather strong sense.

In the context of spam filtering, the group has studied a closely related problem. Each email user receives messages according to an individual, unknown distribution, reflected only in the unlabeled inbox. The spam filter for a user is required to perform well with respect to this distribution. Labeled messages from publicly available sources can be utilized, but they are governed by a distinct distribution, not adequately representing most inboxes. The group has studied how Dirichlet process priors can be employed to transfer knowledge that has been obtained from previous email users to new email accounts.

In conjunction with the ECML-PKDD conference, Steffen Bickel has organized the Discovery Challenge 2006. The challenge provides benchmark data and evaluation protocols for learning from biased data and transfer learning in the email spam filtering application.

Learning from differing training and test distribution and transfer learning remain key challenges for machine learning in the next years, and future research will continue to contribute to this area. Learning problems for which training and test data are governed by differing distributions occur in many areas, and future research will encompass some of these cases in application-oriented collaborations. One such problem is the prediction of the effectiveness of HIV drugs: Since the distribution of distinct versions of the virus changes constantly, past training data no longer represent the current distribution.

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 Ulf Brefeld, Christoph Büscher, and Tobias Scheffer on semi-supervised sequence learning has been distinguished with the Best Paper Award of the European Conference on Machine Learning. Follow-up work has led to a semi-supervised discriminative parsing algorithm for natural language. Recent studies extend the framework to the problem of supervised clustering of streaming data. 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.

Discriminative structural learning methods have a great potential for contributing to better solutions for several problems in bioinformatics, including the prediction of protein interactions. Such problems will be addressed in future collaborations.

Learning for Information Retrieval

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.

Publication repositories contain an abundance of information about the evolution of scientific research areas. Recent work 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, probabilistic, generative models characterize the dependency between the observable documents, and latent factors such the underlying topics and relations between topics.

Similar technologies – generative probabilistic models, Dirichlet process models, MCMC sampling algorithms – are employed to detect hidden causes in workshop reports and warranty databases in cooperation with DaimlerChrysler.

The group studies machine learning technologies that help to obtain a better, personalized relevance ranking of information items. Information about the relevance of information items for individuals can be obtained from *clickthrough data*; *e.g.,* from logs of clicks to result pages of search engines, or logs of clicks to online advertisements, or clicks to products at e-commerce websites.

Adversarial Learning and Security

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 learning problems involve learning about an *adversary* who does attempt to foil the learner. Senders of spam, phishing, and virus emails reverse-engineer any filtering mechanism employed. They change the generators such as to maximize the odds of successfully deceiving it, thus creating a moving target for the learning algorithm. Understanding how learning algorithms can be constructed that anticipate their opponent's next move is a major research challenge. The Machine Learning Groups pursues a collaboration with the European webhosting company STRATO AG.

Projects and Collaborations

In a collaboration with nugg.ad AG, the group investigates on 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. A collaboration with the European web hosting company STRATO AG focuses on a range of security issues and adversarial learning. At STRATO, server intrusions are attempted on a daily basis, and between 30 and 70 million spam, phishing, and virus emails are sent to STRATO's servers. The collaboration has resulted in the development of the STRATO ServerSide security suite.

A collaboration with DaimlerChrysler focuses on the automated discovery of trends and hidden causes in warranty databases and workshop reports. 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 Machine Learning Group intensely collaborates with Departments 5 (Databases and Information Systems) and 3 (Computational Biology) as well as with the Software Engineering Group at Saarland University. Tobias Scheffer is member of the Interdisciplinary Center for Ubiquitous Information of Humboldt-Universität zu Berlin and the Interdisziplinäres Zentrum für Sprachliche Bedeutung at Humboldt-Universität zu Berlin.stems) and 3 (Computational Biology) as well as with the Software Engineering Group at Saarland

University. Tobias Scheffer is member of the Interdisciplinary Center for Ubiquitous Information of Humboldt-Universität zu Berlin and the Interdisziplinäres Zentrum für Sprachliche Bedeutung at Humboldt-Universität zu Berlin.

Group Development

The group was established in January 2007. Steffen Bickel, Ulf Brefeld, Michael Brückner, Laura Dietz, Isabel Drost, Uwe Dick, Peter Haider, and Sascha Schulz left Humboldt-Universität zu Berlin jointly with Tobias Scheffer to join the Max Planck Institute. Thoralf Klein, Barbara Pogorzelska, Christoph Sawade, and Peter Siemen have joined the group as Research Assistants. The group is expected to slowly grow by between one and two members per year.

10 The Discrete Optimization Group (IRG2)

The IRG2 *Discrete Optimization* was established in January 2002 headed by Friedrich Eisenbrand. Since then, he successfully finished his habilitation in December 2003, he was appointed to a professorship (W2) at the university of Dortmund in December 2005, and as of October 2006 he is a full professor (W3) at the university of Paderborn.

Working fields

Our research has been focused around the topics *Integer Programming* and *Combinatorial Optimization*.

Difficult optimization problems are hidden in many different areas of everyday life. In a globalized business, manufacturing and communication world, optimization problems have become large and hard to solve. Mathematical optimization is now a key technology for decision making and the ability to solve difficult optimization problems is a competitive advance.

The progress of optimization software is due to the implementation of ideas and techniques which have been developed by computer scientists and mathematicians.

In respect thereof, we were driven by the following goals:

- Deliver fundamental results in the theory of integer programming and combinatorial optimization, which have high impact in the community.
- Contribute with software to solve difficult applied optimization problems.
- Promote the subject to students by teaching optimization.

History

The IRG2 existed for almost 5 years. Within this time, we achieved many theoretical result. Among them are some of highest impact, e.g. the fastest algorithm for integer programming in the plane. In several projects, e.g. AVACS (see 19.5.1) and the Trunk Packing Project, we supplied software for solving real-world problems. The fastest software currently available for counting and enumerating 0/1 points in a polytope was developed in our group.

In addition to that, we were also dedicated to teaching. We gave several lectures and reading groups. One lecture was voted "Best lecture" by the computer science students of the Saarland University and Friedrich Eisenbrand received the "Teaching Innovation Award" for it. Several master thesises were written in our group.

The PhD students who started in the IRG2, have been seamlessly integrated in D1 or moved to the university of Paderborn. Some of their dissertations will be finished this year. Their cooperations with Friedrich Eisenbrand still exist.

11 The Graphics – Optics – Vision Group (IRG3)

In 2002, the Independent Research Group "Graphics–Optics–Vision" was established at MPII with the goal to solve interdisciplinary challenges from computer graphics, computer vision, and applied optics. In the following years, the group was able to make a number of internationally recognized contributions to the fields of image analysis and synthesis. With the appointment of the head of the group to full professor at TU Braunschweig in January 2006, the group members who wished to stay at MPII joined D4.

Cooperations

From 2005-2007, members of IRG3 went for several-month-long research visits to Microsoft Research Asia in Bejing, China, and to the University of British Columbia in Vancouver, Canada. Research cooperations with resulting joint publications were also maintained with ETH Zurich, University of Aberdeen, Technion Haifa, INRIA Sophia-Antipolis, and Weimar University. The group was an active member of the EU Network of Excellence "3D-TV". The newly established group at TU Braunschweig keeps cordial ties to MPII and pursues several research projects in close cooperation with the Computer Graphics Department D4.

Group Development

Christian Linz joined the group in October 2005, adding to the group his expertise in image-based reconstruction. In January 2006, the head of the group, Marcus Magnor, left MPII to take up his new position as full professor at TU Braunschweig. With Timo Stich and Christian Linz, two junior PhD students decided to accompany Marcus Magnor to TU Braunschweig in Spring 2006. The four senior PhD students Lukas Ahrenberg, Ivo Ihrke, Andrei Lintu, Volker Scholz decided to stay at MPII and finish their research under the supervision of Hans-Peter Seidel. In July 2006, the group's first PhD student, Bastian Goldlücke, received his doctoral degree "with distinction".

Part III

Research Units in Detail

12 The Algorithms and Complexity Group (D1)

12.1 Personnel

Director

Prof. Dr. Kurt Mehlhorn

Current Researchers

Dr. Ernst Althaus
Dr. Holger Bast
Dr. René Beier
Dr. Kevin Chang
Dr. Giorgos Christodoulou
Dr. Benjamin Doerr
Dr. Khaled Elbassioni
Dr. Stefan Funke
Dr. Anders Gidenstam
Dr. Joachim Giesen
Dr. Sathish Govindarajan
Dr. Nils Hebbinghaus
Dr. Ulrich Meyer
Dr. Dimitrios Michail
Dr. Frank Neumann
Dr. Alantha Newman
Dr. Michael Sagraloff

Former (Recent) Researchers

Dr. Surender Baswana (–June 2006, now Asst. Prof. at IIT Kanpur)
Dr. Aleksei Fishkin (–August 2005, now at Siemens)
Dr. Gianni Franceschini (–February 2005, now at Univ. of Pisa)
Dr. Lutz Kettner (–December 2005, now at Mental Images, Berlin)
Dr. Lukasz Kowalik (–September 2006, now at Univ. of Warsaw)
Dr. Martin Kutz
Dr. Marcin Mucha (–September 2006, now at Univ. of Warsaw)
Dr. Nabil Mustafa (–September 2006, now Asst. Prof. at Lahore Univ. of Management Sciences)
Dr. Katarzyna Paluch (–September 2006, now at Univ. of Wroclaw)

Dr. Seth Pettie (–June 2006, now Asst. Prof. at Univ. of Michigan)
Dr. René Sitters (–October 2006, now at Univ. of Eindhoven)
Dr. Nicola Wolpert (–September 2005, now Asst. Prof. at Hochschule für Technik Stuttgart)

PhD Students

Deepak Ajwani
Markus Behle
Eric Berberich
Stefan Canzar
Arno Eigenwillig
Tobias Friedrich
Edda Happ
Daniel Johannsen
Kanela Kaligosi
Andreas Karrenbauer
Michael Kerber
Christian Klein
Annamaria Kovacs
Domagoj Matijevic
Andreas Meyer
Rouven Naujoks
Ralf Osbild
Vitaly Osipov
Evangelia Pyrga
Imran Rauf
Pascal Schweitzer
Ingmar Weber

Secretaries

Ingrid Finkler
Petra Mayer

12.2 Visitors

In the time period from June 2005 to March 2007, the following researchers visited our group:

Marcin Mucha	06.04.05–05.05.05	University of Warsaw
Tomasz Walen	06.04.05–05.05.05	University of Warsaw
Peter Korteweg	25.04.05–29.04.05	Eindhoven Univ. of Tech.
Uri Zwick	10.05.05–15.05.05	Tel Aviv University
N.S. Narayanaswamy	13.05.05–29.07.05	Indian Inst. of Tech. Madras
Mihai Patrascu	12.06.06–16.06.05	MIT
Kavitha Telikepalli	16.06.05–31.07.05	Indian Inst. of Sci. Bangalore
Alexander Kononov	24.05.05.–25.05.05	Novosibirsk Sob. Inst. of Math.

Sergei Sevastianov	29.05.05–05.06.05	Novosibirsk Sob. Inst. of Math.
Yan Zhang	28.06.05–20.08.05	University of Hong Kong
Zhen Zhou	28.06.05–02.09.05	University of Hong Kong
Alantha Newman	16.06.05–17.06.05	RWTH Aachen
Rudolf Fleischer	20.07.05–23.07.05	Fudan University Shanghai
Uri Zwick	28.08.05–02.09.05	Tel Aviv University
Rolf Klein	28.08.05–02.09.05	Universität Bonn
Günter Rote	28.08.–02.09.05	Freie Universität Berlin
Jeff Erickson	28.08.05–03.09.05	Univ. of Illinois at Urbana-Champaign
Tobias Baumann	13.09.05–14.09.05	Daimler-Chrysler
Zvi Lotker	16.09.05–22.09.05	CWI Amsterdam
Nico Kruithof	24.09.05–01.10.05	University of Groningen
Olga Gerber	10.10.05–11.10.05	Universität Kiel
Vikram Sharma	09.10.05–17.10.05	New York University
Aleksander Madry	09.11.05–31.12.05	University of Warschau
Frank Neumann	14.11.05–21.11.05	Universität Kiel
Nicola Milosavljevic	01.11.05–31.01.06	Stanford University
Michal Mayerovitch	04.12.05–22.12.05	University of Tel Aviv
Christian Sohler	26.01.06–31.01.06	University of Paderborn
Alex Fishkin	05.02.06–18.02.06	University of Liverpool
Lars Arge	03.03.06–17.03.06	Duke University
Norbert Zeh	03.03.06–17.03.06	Dalhousie University Halifax
Pashant Batra	20.02.06–24.02.06	University Hamburg-Harburg
Tobias Storch	03.05.06–04.05.06	Universität Dortmund
Carsten Witt	03.05.05–04.05.06	Universität Dortmund
Frank Neumann	03.05.06–05.05.06	Universität Kiel
Kavitha Telikepalli	15.05.06–30.07.06	Indian Inst. of Sci. Bangalore
Anand Bhalgat	11.06.06–30.07.06	Indian Inst. of Sci. Bangalore
Tetsuo Asano	09.06.06–30.08.06	Japan Advanced Inst. of Sci.
Naveen Garg	01.06.06–31.05.07	Indian Inst. of Tech. Delhi
Amit Kumar	29.05.06–20.07.06	Indian Inst. of Tech. Delhi
Jian Li	03.07.06–21.07.06	Fudan University Shanghai
Qin Zhang	03.07.06–21.07.06	Fudan University Shanghai
Zhang, Yan	10.07.06–10.09.06	University of Hong Kong
Zvi Lotker	10.07.06–30.09.06	CWI, Amsterdam
Guy Even	04.08.06–18.08.06	Tel Aviv University
Rohit Khandekar	22.08.06–23.09.06	University of Waterloo
Jochen Koenemann	17.08.06–24.08.06	University of Waterloo
Kazuhisa Makino	01.09.06–10.09.06	Univerity of Tokyo
Ingo Wegener	21.08.06–25.08.06	Universität Dortmund
Tamal Dey	21.08.06–25.08.06	Ohio State University
Joel Spencer	21.08.06–26.08.06	New York University
Andreas Beckmann	03.09.06–09.09.06	Universität Halle

Heiko Röglin	24.09.06–27.09.06	RWTH Aachen
Meena Mahajan	01.11.06–07.11.06	CIT Campus Chennai India
Alexander Fanghäusel	15.10.06–17.10.06	Universität Jena
Tanja Gernhard	05.11.06–10.11.06	Universität München
Martin Skutella	16.11.06	Universität Dortmund
Tobias Storch	25.11.06–27.11.06	Universität Dortmund
Eckart Zitzler	27.11.06–29.11.06	ETH Zürich
Dimo Brockhoff	27.11.06–29.11.06	ETH Zürich
Renato Renner	27.11.06–29.11.06	Cambridge University
Rob van Stee	12.01.07	Universität Karlsruhe
Torben Hagerup	22.02.07–06.04.07	Universität Augsburg
Nicola Wolpert	08.03.07–09.03.07	Hochschule f. Technik Stuttgart
Thomas Sauerwald	12.03.07–14.03.07	Universität Paderborn
Jie Gao	03.04.07–06.04.07	New York University at Stony Brook

Group Organization

We have different kind of meetings:

- We have a Noon Seminar, in which we present our own work as well as the work of others. The noon seminar is also used for guest speakers. It takes place once or twice a week on average.
- We have reading groups on approximation algorithms (Naveen Garg and Khaled Elbassioni), discrete mathematics (Benjamin Doerr), and curves and algorithms (Michael Sagraloff and Oliver Labs(Math Department))
- The senior researchers (steering committee) of D1 have a formal meeting at least monthly and informal meetings almost daily.

12.3 Foundations and Discrete Mathematics

Coordinator: Benjamin Doerr

The research area "Foundations and Discrete Mathematics" has its roots in the area "Foundations", existing since 2000 and previously coordinated by Seth Pettie, and the area "Discrete Mathematics", which was founded in 2005 and coordinated by Benjamin Doerr since then. After Seth Pettie took up a professorship in Michigan, U.S.A, in 2006, the two areas were joined to one.

Our focus is to investigate those structures and principles that are fundamental for the understanding and design of efficient algorithms. This includes mathematical objects like graphs and hypergraphs, methods from probability theory and classical computer science areas like complexity theory. Since these ingredients are crucial for many, also more applied problems, there is a rich exchange with the other research areas in D1.

The particular topics we work on are subject to change, influenced both by people joining and leaving the group and by current developments in algorithmics. Topics that emerged or got much more attention in the last two years include discrepancy theory, theoretical analyses of evolutionary algorithms and mechanism design; classical topics are graph theory and algorithms, data structures and computer algebra.

12.3.1 Discrepancy Theory

Discrepancy theory deals with a number of different, though intimately related problems. One common theme is *approximating* something large, complicated or continuous by a smaller, simpler or discrete object in a way that the characteristic properties are maintained to a sufficient extent. Often we aim at maintaining several properties simultaneously. In this case, we try to keep a fair balance between them. Hence *fairness* and *balance* is another central theme in discrepancy theory.

Discrepancy theory originates from number theory. In the early years of the 20th century, mathematicians noted that is it highly non-trivial to construct sequences of numbers in the unit interval $[0,1]$ such that all subintervals $[a,b] \subseteq [0,1]$ for all $n \in \mathbb{N}$ contain roughly $(b-a)n$ of the first n terms of the sequence. However, such *uniformly distributed* sequences were found to be very useful in numerical integration. It took several decades and deep results like [3, 4, 1, 2, 5] to fully understand the problem and see that the *discrepancy* here is of order $\Theta(\log n)$. By this we mean the following: On the one hand, you cannot find a perfectly distributed sequence. For any sequence in $[0,1]$, there are $a, b \in [0,1]$ and $k \leq n$ such that among the first k terms of the sequence, not $(b-a)k$ lie in the interval $[a,b]$, but some number deviating from this by at least $\Omega(\log n)$. On the other hand, there are sequences such that for all n, $k \leq n$ and $a, b \in [0,1]$ these deviations are at most of order $\log n$.

From this fundamental insight that "complete (dis)order is impossible" the area of discrepancy theory developed. Similar phenomena are found in several branches of mathematics and computer science. As an algorithmic example, let us look at rounding problems.

Rounding arbitrary numbers to integers is a key problem in optimization. One reason for this is that continuous problems often can be solved more easily than discrete ones. When rounding, we naturally want to maintain several properties of the original numbers, e.g., that they satisfy several linear equations. Discrepancy theory tells us that such rounding usually is not possible in the perfect sense, but does work if we allow some discrepancies. More precisely, you can have n numbers (say in form of a vector $x \in [0,1]^n$) and n linear equations (encoded in an $n \times n$ matrix A, say normalized such that all $|a_{ij}| \leq 1$) such that any rounding $y \in \{0,1\}^n$ of x yields a discrepancy of order $\Omega(\sqrt{n})$ in some equation, that is, $\|(Ax) - (Ay)\|_\infty = \Omega(\sqrt{n})$. Even worse, in a certain sense most x and A display such a behavior. On the positive side, if we are willing to tolerate some discrepancies, things can be very easy. Randomized rounding (see also further down this subsection) is a simple way to find a rounding y such that $\|(Ax) - (Ay)\|_\infty = O(\sqrt{n \log n})$.

In the sections below, we briefly describe the several discrepancy problems we worked on over the last two years.

References

[1] T. van Aardenne-Ehrenfest. Proof of the impossibility of a just distribution of an infinite sequence of points. *Nederl. Akad. Wet., Proc.*, 48:266–271, 1945.

[2] T. van Aardenne-Ehrenfest. On the impossibility of a just distribution. *Nederl. Akad. Wet., Proc.*, 52:734–739, 1949.

[3] J. G. van der Corput. Verteilungsfunktionen I. *Akad. Wetensch. Amsterdam, Proc.*, 38:813–821, 1935.

[4] J. G. van der Corput. Verteilungsfunktionen II. *Akad. Wetensch. Amsterdam, Proc.*, 38:1058–1066, 1935.

[5] W. M. Schmidt. On irregularities of distribution VII. *Acta Arith.*, 21:45–50, 1972.

Randomized Rounding and Derandomization

Investigator: Benjamin Doerr

Many combinatorial optimization problems can easily be formulated as integer linear programs (ILPs). Unfortunately, solving ILPs is NP–hard, whereas solving linear programs (without integrality constraints) is easy, both in theory and practice. Therefore, a natural and widely used technique is to solve the linear relaxation of the ILP and then transform its solution into an integer one.

Typically, this requires rounding a vector x to an integer one y in such a way that the rounding errors $|(Ax)_i - (Ay)_i|$, $i \in [m] := \{1, \dots, m\}$, are small for some given $m \times n$ matrix A.

A very successful approach to such rounding problems (and many others) is the one of *randomized rounding* introduced by Raghavan and Thompson [6, 5]. Here the integer vector y is obtained from x by rounding each component j independently with probabilities derived from the fractional part of x_j. In particular, if $x \in [0,1]^n$, we have $\Pr(y_j = 1) = x_j$ and $\Pr(y_j = 0) = 1 - x_j$ independently for all $j \in [n]$.

Since the components are rounded independently, the rounding error $|(Ax)_i - (Ay)_i|$ in constraint i is a weighted sum of independent random variables. Thus it is highly concentrated around its mean, which by choice of the probabilities is zero. Large deviation bounds like the Chernoff inequality allow to quantify such violations and thus yield performance guarantees.

The *derandomization* problem is to transform this randomized approach into a deterministic rounding algorithm that keeps the rounding errors $|(Ax)_i - (Ay)_i|$ in a similar range as randomized rounding would have done with positive probability.

In our work, we extended the classical randomized rounding approach of Raghavan and Thompson to include certain hard constraints, where we do not tolerate any violations. A second result is a very simple way to derandomize randomized rounding in the general case of rational input numbers, which was a long-standing open problem.

Generating Randomized Rounding with Cardinality Constraints. A *cardinality constraint* is a constraint prescribing that the sum of certain variables shall have a particular value. For binary variables, this means that a particular number of them shall be one, which explains

the term "cardinality constraint". Whereas independent randomized rounding is very good in ensuring small rounding errors, it can not guarantee that a cardinality constraint remains unviolated. This is a real problem, as many optimization problem do include cardinality constraints, and these are often too important to allow even slight violations.

Since independent randomized rounding does not respect such hard cardinality constraints, we need a way to generate randomized roundings that automatically satisfy such constraints. Naturally, this means that this way of rounding has to be dependent. However, we would still like it to be almost independent in the sense that we still have large deviations bounds.

Apart from few simple cases, this problem is highly non-trivial. Hence the first solution for such a problem, due to Srinivasan [7], was only presented at FOCS 2001. It allows a single cardinality constraint involving all variables. It works in linear time (as independent randomized rounding), but is quite technical. As a consequence, Srinivasan did not give a derandomization, and it seems unlikely that this approach can be derandomized. In [1], we gave an alternative way to generate such randomized roundings. It seems less complicated, at least, it can be derandomized easily with the standard methods for the independent case.

A year later, Srinivasans pioneering work was extended [4] to the so-called bipartite edges weight rounding problem, which is a rounding problem in which each variable may occur in at most two cardinality constraints. Here, rounding in linear time was not possible anymore, but the run-time had the number of hard constraints as extra factor. Again, a derandomization was not in sight. This was overcome again in [2, 3]. The precise results are too complicated to state here in detail, but roughly speaking, we showed the following. We give an alternative way to generate randomized roundings for the bipartite edge weight rounding problem. It runs in linear time in the bit-model for input numbers having a finite binary expansion. It also works for arbitrary rational numbers, still in the more robust bit-model, but the run-time bounds are pseudo-polynomial if the denominators involve large prime factors. Derandomization is possible and efficient in all cases.

The heart of all three results is a deep theoretical insight. We showed that randomized rounding with hard cardinality constraints and negative correlation (which imply Chernoff bounds) and its derandomization is possible for all input vectors from $\mathbb{Q}^n$ if and only if it is possible for all input vectors from $\{0, \frac{1}{2}\}^n$. This reduction is constructive, that is, if one can solve the half-integral case, one immediately has an efficient solution for the general one. From the view-point of algorithm design, this means that only the half-integral case has to be solved, which often is much easier. In addition, often this half-integral case can be solved more efficiently. The above-mentioned run-time improvements are due to this fact.

Derandomizing Randomized Rounding. Using the binary expansion idea again, we improve classical results on derandomizing randomized rounding (without constraints, though constraints could be added using the ideas sketched above). In particular, we give an $O(mn \log n)$ time derandomization for arbitrary constraint matrices $A \in ([0,1] \cap \mathbb{Q})^{m \times n}$ that works in the RAM model. This improves over the $O(mn^2 \log(mn))$ solution of Srivastav and Stangier [8]. Note that Raghavan's derandomization [5] needs to compute the exponential function and in consequence in the RAM model only works for binary matrices. This is pointed out in Section 2.2 of his paper, but seemingly overlooked by many who cite it.

A second advantage of our derandomization is its simplicity. We solve the problem by $\log n$

times invoking Raghavan's derandomization for some binary matrix. This is considerably simpler than approximating the exponential function in the large deviation inequalities via Taylor polynomials as in [8].

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. Randomly rounding rationals with cardinality constraints and derandomizations. In W. Thomas and P. Weil, eds., *STACS 2007*, Aachen, 2007, *LNCS 4393*, pp. 441–452. Springer.

[4] R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan. Dependent rounding in bipartite graphs. In *Proc. 43rd Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*, 2002, pp. 323–332.

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

[6] P. Raghavan and C. D. Thompson. Randomized rounding: A technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7:365–374, 1987.

[7] 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.

[8] A. Srivastav and P. Stangier. Algorithmic Chernoff–Hoeffding inequalities in integer programming. *Random Structures and Algorithms*, 8:27–58, 1996.

Matrix Rounding and Applications

Investigators: Benjamin Doerr, Tobias Friedrich, Christian Klein, and Ralf Osbild

Rounding a real-valued matrix to an integer one such that the rounding errors in all rows and columns are less than one is a classical problem. It has been applied to hypergraph coloring, in scheduling and in statistics. We have developed several algorithms to compute such roundings.

Scheduling In the *flexible transfer line scheduling problem* the goal is to produce m different goods on a single machine in a balanced manner. The demands $d_i \in \mathcal{N}$, $i \in [1..m]$ for each good are known in advance. Under the assumption that each good can be produced in one unit of time and that there are no switch-over costs between products, this problem can be formulated as a matrix rounding problem. Each row of the matrix represents one kind of goods and each column represents the demand at a given time step. A good schedule is then represented by a rounding that has rounding error less than one in all columns and in all initial intervals of rows, i.e. the first i elements of a given row. We give a linear-time one-pass algorithm to solve this rounding problem in [1].

Statistics Another application of matrix-rounding is to process a table of statistical data prior to its publication. Here the entries of the table should be rounded to multiples of a given base (e.g., multiples of 10). Such a rounding can be used to improve the readability of data tables. The main reason, however, to apply such a rounding procedure is confidentiality protection. Frequency counts that directly or indirectly disclose small counts may permit the identification of individual respondents, hence one wants to round such entries to a multiple of a given base.

For such statistical tables it is obviously also desirable that the row and column sums are not changed by the rounding procedure. This is known as *Controlled Rounding* in statistics.

We developed several algorithms that compute controlled roundings. They all satisfy the additional constraints that the rounding error in all initial intervals of rows and columns is less than one. This is particular useful when there is a linear ordering on statistical attributes.

If the data is to be processed further, a so-called *unbiased rounding* may be desirable. This means that the rounding is computed by a randomized algorithm in such a way that the expected value of each rounded entry equals its original value.

In [2] we present an algorithm that computes a controlled rounding satisfying the above mentioned additional constraints in time $O(n \log n)$. The algorithm can be modified to compute unbiased roundings. For this, however, the entries of the input matrix must have finite bit-length.

For base 10 however, the matrix entries are rational numbers with infinite bit-length, hence the above algorithm cannot be applied. In [3] we present an algorithm that can handle rational $m \times n$-matrices in expected time $O(mnq^2)$, where q is the common denominator of the matrix entries. We also show that if the denominator can be written as $q = \prod_{i=1}^{\ell} q_i$ for some integers q_i, the expected runtime can be reduced to $O(mn \sum_{i=1}^{\ell} q_i^2)$. A derandomization of the algorithm is also discussed.

References

- [1] B. Doerr, T. Friedrich, C. Klein, and R. Osbild. Rounding of sequences and matrices, with applications. In T. Erlebach and P. Persiano, eds., *Third Workshop on Approximation and Online Algorithms (WAOA 2005)*, Palma de Mallorca, Spain, 2005, *LNCS 3879*, pp. 96–109. Springer.
- [2] B. Doerr, T. Friedrich, C. Klein, and R. Osbild. Unbiased matrix rounding. In L. Arge and R. Freivalds, eds., *Algorithm theory - SWAT 2006, 10th Scandinavian Workshop on Algorithm Theory*, Riga, Latvia, 2006, *LNCS 4059*, pp. 102–112. Springer.
- [3] B. Doerr and C. Klein. Unbiased rounding of rational matrices. In S. Arun-Kumar and N. Garg, eds., *FSTTCS 2006: Foundations of Software Technology and Theoretical Computer Science: 26th International Conference*, Kolkata, India, 2006, *LNCS 4337*, pp. 200–211. Springer.

Constructing Uniformly Distributed Point Sets

Investigators: Benjamin Doerr in collaboration with Michael Gnewuch (Kiel) and Anand Srivastav (Kiel)

The L_∞– or *star-discrepancy* of an n–point set P in the unit cube $[0,1]^d$ is $\mathrm{disc}(P) = \sup_R |\frac{1}{n}|P \cap R| - \lambda(R)|$, where λ denotes the Lebesgue measure and R ranges over all axis-parallel rectangles containing the origin (as "lower left corner"). Point sets with low dis-

crepancy are important for numerical integration. The error inflicted by taking the average function value instead of the true integral is bounded by the product of the discrepancy of the sample point set and the variation (in the sense of Hardy and Krause) of the function:

$$\left| \frac{1}{n} \sum_{p \in P} f(p) - \int_{[0,1]^d} f(x)dx \right| \leq \mathrm{disc}(P)V(f).$$

This is the so-called Koksma-Hlawka inequality [8, 7]. Since we cannot change the function we need to integrate, we can only reduce the integration error by choosing low-discrepancy point sets.

From the theoretical point of view, the problem is reasonably well understood. There are n–point sets having discrepancy $O(\frac{1}{n}\log^{d-1} n)$, and there cannot be ones having a discrepancy of less than $\Omega(\frac{1}{n}\log^{(d-1)/2} n)$. These are famous deep results due to van der Corput [1, 2], Halton [5], Hammersley [6] and Roth [9]. As famous is "the big open problem of geometric discrepancies" of closing the gap between these bounds.

Unfortunately, for the practical application in reasonable high dimension – practitioners typically quote values between 20 and 360 – these bounds are utterly useless. Note that for $n < e^{d-1}$, $\frac{1}{n}\log^{d-1}(n)$ is an increasing function! In addition, the implicit constants in the bounds above depend terribly on d. Therefore, in these cases the current best point sets are obtained from randomized methods.

Due to their high technicality, we can only roughly outline our results here. The interested reader is invited to read the original papers. In [4], we use randomized rounding and its derandomization to compute low-discrepancy point sets. To make the whole problem a rounding problem, we partition the unit cube into suitable sub-regions ("a so-called delta-cover") and aim at having the right number of points in all these regions. Since these "right numbers of points" are usually fractional, we end up with a rounding problem. This approach is refined in [3], where instead of independently rounding we use the dependent rounding approach described in Section 12.3.1. This allows to greatly reduce the number of variables to be rounded.

References

[1] J. G. van der Corput. Verteilungsfunktionen I. *Akad. Wetensch. Amsterdam, Proc.*, 38:813–821, 1935.

[2] J. G. van der Corput. Verteilungsfunktionen II. *Akad. Wetensch. Amsterdam, Proc.*, 38:1058–1066, 1935.

- [3] B. Doerr and M. Gnewuch. Construction of low-discrepancy point sets of small size by bracketing covers and dependent randomized rounding. Research Report 06-14, University Kiel, Kiel, 2006.
- [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, October 2005.

[5] J. H. Halton. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numer. Math.*, 2:84–90, 1960.

[6] J. M. Hammersley. Monte Carlo methods for solving multivariable problems. *Ann. New York Acad. Sci.*, 86:844–874, 1960.

[7] E. Hlawka. Funktionen von beschränkter Variation in der Theorie der Gleichverteilung. *Ann. Math. Pura Appl.*, 54:325–333, 1961.

[8] J. F. Koksma. A general theorem from the uniform distribution modulo 1 (in dutch). *Mathematica B (Zutphen)*, 1:7–11, 1942/43.

[9] K. F. Roth. On irregularities of distribution. *Mathematika*, 1:73–79, 1954.

Hypergraph Coloring

Investigators: Benjamin Doerr, Mahmoud Fouz, Nils Hebbinghaus, and Daniel Johannsen in collaboration with Michael Gnewuch (Kiel) and Sören Werth (Kiel)

The combinatorial discrepancy problem is to partition the vertices of a hypergraph in such a way that this partition induces an even partition on all hyperedges. The discrepancy of a hypergraphs measures how well this aim can be achieved. This notion is both fundamental in discrete mathematics and related to applications as rounding and declustering.

Hereditary Discrepancy. The hereditary discrepancy of a hypergraph is the maximal discrepancy of all its induced sub-hypergraphs. In [7, 3] we bounded the hereditary discrepancy of a hypergraph $\mathcal{H}$ in two colors in terms of its hereditary discrepancy in c colors. We showed that $\mathrm{herdisc}(\mathcal{H}, 2) \leq Kc\,\mathrm{herdisc}(\mathcal{H}, c)$, where K is some absolute constant. This bound is sharp.

Linear Discrepancy The *linear discrepancy* or *lattice approximation problem* is to round the real valued solution x of a linear system of equations $Ax = b$ to an integer valued solution z such that $||Az - b||_\infty$ is minimized. The maximum (over all x) of these minima is the linear discrepancy of the matrix A. This notion extends to hypergraphs via the incidence matrix.

A classical result [9] states that the linear discrepancy of a hypergraph can be bounded in terms of its hereditary discrepancy. However, this upper bound is not sharp. Only for unimodular hypergraphs this problem is fully understood [2]. For general matrices, the current best bounds are from [1]. In the ongoing study of the relation between linear and hereditary discrepancy, we produced simpler proofs of some known results, but aim at improving the existing bounds.

Arithmetic Progressions. Estimating the discrepancy of the hypergraph of all arithmetic progressions in the set $[N] = \{1, 2, \ldots, N\}$ was one of the famous open problems in combinatorial discrepancy theory for a long time. A generalization of this classical hypergraph is the hypergraph $\mathcal{H}$ of sums of two arithmetic progressions. The hyperedges of this hypergraph are of the form $A_1 + A_2$ in $[N]$, where the A_i are arithmetic progressions. The probabilistic method gives an upper bound of order $O((N \log N)^{1/2})$. In [8] we proved a lower bound of order $\Omega(N^{1/2})$ the discrepancy $\mathrm{disc}(\mathcal{H})$.

Symmetric Products of Hypergraphs. For a hypergraph $\mathcal{H} = (V, \mathcal{E})$, its d–fold symmetric product is $\Delta^d\mathcal{H} = (V^d, \{E^d | E \in \mathcal{E}\})$. In [4] we gave several upper and lower bounds for the c–color discrepancy of such products. In particular, we showed that the bound $\mathrm{disc}(\Delta^d\mathcal{H}, 2) \leq \mathrm{disc}(\mathcal{H}, 2)$ proven for all d by Doerr, Srivastav and Wehr [6] cannot be extended to more than

$c = 2$ colors. In fact, for any c and d such that c does not divide $d!$, there are hypergraphs having arbitrary large discrepancy and $\mathrm{disc}(\Delta^d\mathcal{H}, c) = \Omega_d(\mathrm{disc}(\mathcal{H}, c)^d)$. Apart from constant factors (depending on c and d), in these cases the symmetric product behaves no better than the general direct product $\mathcal{H}^d$, which satisfies $\mathrm{disc}(\mathcal{H}^d, c) = O_{c,d}(\mathrm{disc}(\mathcal{H}, c)^d)$.

Declustering Problem. The declustering problem is to allocate given data on parallel working storage devices in such a manner that typical requests find their data evenly distributed on the devices. Using deep results from discrepancy theory, we improved in [5] previous work of several authors concerning range queries to higher-dimensional data. We gave a declustering scheme with an additive error of $O_d(\log^{d-1} M)$ independent of the data size, where d is the dimension, M the number of storage devices and $d-1$ does not exceed the smallest prime power in the canonical decomposition of M into prime powers. In particular, our schemes work for arbitrary M in dimensions two and three. For general d, they work for all $M \geq d-1$ that are powers of two. Concerning lower bounds, we showed that a recent proof of a $\Omega_d(\log^{\frac{d-1}{2}} M)$ bound contains an error. We closed the gap in the proof and thus established the bound. One direction of the ongoing research is to generalize the declustering problem to parallel storage devices running with different speed.

References

[1] B. Doerr. Linear and hereditary discrepancy. *Combinatorics, Probability and Computing*, 9:349–354, 2000.

[2] B. Doerr. Linear discrepancy of totally unimodular matrices. *Combinatorica*, 24:117–125, 2004.

- [3] B. Doerr and M. Fouz. Hereditary discrepancies in different numbers of colors ii, November 2006.
- [4] B. Doerr, M. Gnewuch, and N. Hebbinghaus. Discrepancy of symmetric products of hypergraphs. *The Electronic Journal of Combinatorics*, 13:1–12, 2006.
- [5] B. Doerr, N. Hebbinghaus, and S. Werth. Improved bounds and schemes for the declustering problem. *Theoretical Computer Science*, 359(1-3):123–132, 2006.

[6] B. Doerr, A. Srivastav, and P. Wehr. Discrepancy of Cartesian products of arithmetic progressions. *Electron. J. Combin.*, 11:Research Paper 5, 16 pp. (electronic), 2004.

- [7] M. Fouz. Hereditary discrepancy in different numbers of colors. Bachelor thesis, Universität des Saarlandes, October 2006.
- [8] N. Hebbinghaus. Discrepancy of sums of two arithmetic progressions. ArXiv math.NT/0703108, Cornell University (ArXiv), ArXiv, March 2007.

[9] L. Lovász, J. Spencer, and K. Vesztergombi. Discrepancies of set-systems and matrices. *Europ. J. Combin.*, 7:151–160, 1986.

Liar Games

Investigators: Benjamin Doerr in collaboration with Johannes Lengler (Universität des Saarlandes) and David Steurer (Princeton)

The general rules of a liar game are as follows: There are two players, Carole and Paul. Carole chooses a number from $\{1, \ldots, n\}$. There are q rounds. Each round, Paul asks a

Yes/No-question, which Carole answers. In doing so, Carole may lie according to further specifications. Paul wins the game, if after q such rounds, he knows the number.

Liar games are widely used to model online-recovery of communication errors. For this reason, a number of analyses of different liar games exist in the computer science literature. The most important task is to determine the minimal number of questions Paul needs, and to describe an optimal strategy for Paul.

In [1], we analyzed the Burst Error Liar Game, in which all lies (errors) must be contained in some small time interval. This game is motivated by the idea that in a noisy channel, transmission errors will usually not occur independently, but rather as a single, short-time outburst of errors. We showed that in this setting, error correction can be done much more efficiently than in the case of independently distributed errors. In fact, correction can asymptotically be done as efficiently as in the case where only one single bit is affected.

Up to this point, each liar games had to be investigated individually. Only occasionally, it was possible to analyze a class of liar games in a single approach. However, in a forthcoming paper we provided a general method to model and analyze liar games. We assigned to each liar game a "formal" game, which could easily be analyzed by means from the theory of automata and languages. For a class of games, including all games where the total number of lies is bounded independent of the number of rounds, we showed how to compute the solution of the liar game from the solution of the formal game for all but finitely many values of n. In particular, our results hold for most games that were investigated so far in the literature. Our result is more precise than most of the results that were obtained in more specific situations.

References

- [1] B. Doerr, J. Lengler, and D. Steurer. The interval liar game. In T. Asano, ed., *Algorithms and Computation, 17th International Symposium, ISAAC 2006*, Kolkata, India, 2006, *LNCS 4288*, pp. 318–327. Springer.

The Rotor-Router Model (Propp-Machine)

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

Jim 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. We have investigated how well this process simulates a random walk. On grids $\mathbb{Z}^d$ it is known that, apart from a technicality, independent of the starting configuration, at each time, the number of chips on each vertex in the Propp model deviates from the expected number of chips in the random walk model by at most a constant c_d [3].

For the graph being the infinite path, we have shown that this constant c_1 is approximately 2.29 [1, 2]. For intervals of length L, this improves to a difference of $O(\log L)$, for the L_2 average of a contiguous set of intervals even to $O(\sqrt{\log L})$. We also analyzed the difference between Propp machine and random walk on the infinite two-dimensional grid [5, 6]. We have shown that there the constant c_2 is approximately 7.83, if all vertices serve their neighbors in

clockwise or counterclockwise order and 7.29 otherwise. This result in particular shows that the order in which the neighbors are served makes a difference. Our analysis also reveals a number of further unexpected properties of the two-dimensional Propp machine.

Additionally to above discrepancy results, the Propp machine has also been examined in an aggregating model called Internal Diffusion-Limited Aggregation (IDLA) [4]. There, each chip starts at the origin of $\mathbb{Z}^d$ and walks till it reaches an unoccupied site, which it then occupies. In the random walk model it is well known that the shape of the occupied locations converges to a Euclidean ball in $\mathbb{R}^d$ [7]. Recently, Levine and Peres [8] proved an analogous result for the two-dimensional Propp machine. Surprisingly, the convergence seems to be much faster. The figure shows the Propp aggregation with one million particles. The colors denote the final rotor directions.

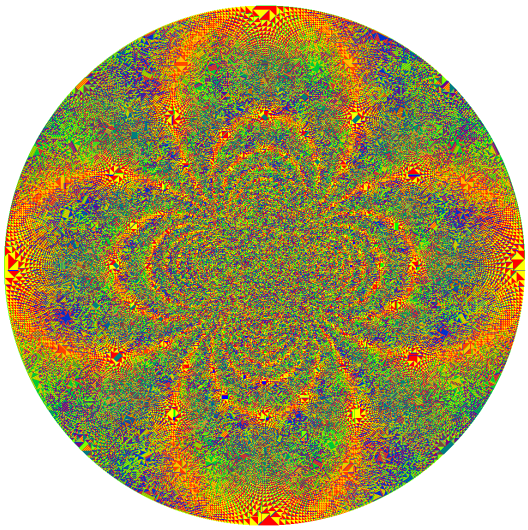


Figure 12.1: Aggregation of one million chips in the Propp model. The colors denote the final rotor directions.

References

- [1] J. Cooper, B. Doerr, J. Spencer, and G. Tardos. Deterministic random walks. In R. Raman, R. Sedwgewick, and M. F. Stallmann, eds., *Proceedings of the Eighth Workshop on Algorithm Engineering and Experiments and the Third Workshop on Analytic Algorithmics and Combinatorics (ALENEX'06 / ANALCO'06)*, Miami, FL, USA, 2006, pp. 185–197. SIAM.
- [2] J. Cooper, B. Doerr, J. Spencer, and G. Tardos. Deterministic random walks on the integers, 2006.

[3] J. Cooper and J. Spencer. Simulating a random walk with constant error. *Combinatorics, Probability and Computing*, 2004. To appear, preliminary version available from arXiv:math/0402323.

[4] P. Diaconis and W. Fulton. A growth model, a game, an algebra, Lagrange inversion, and characteristic classes. *Rend. Sem. Mat. Univ. Pol. Torino*, 49(1):95–119, 1990.

- [5] 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, December 2006, *LNCS 4288*, pp. 474–483. Springer.

- [6] B. Doerr and T. Friedrich. Deterministic random walks on the two-dimensional grid. Submitted to Combinatorics, Probability and Computing, preliminary version available from arXiv:math/0703453, 2007.

[7] G. F. Lawler, M. Bramson, and D. Griffeath. Internal diffusion limited aggregation. *Annals of Probability*, 20(4):2117–2140, 1992.

[8] L. Levine and Y. Peres. The rotor-router shape is spherical. *The Mathematical Intelligencer*, 27(3):9–11, 2005.

12.3.2 Evolutionary Computation

Investigators: Benjamin Doerr, Tobias Friedrich, Nils Hebbinghaus, Daniel Johannsen, Christian Klein, and Frank Neumann

Evolutionary computation methods such as evolutionary algorithms or ant colony optimization have widely been applied to complex engineering problems as well as to problems from combinatorial optimization. We investigate these randomized search heuristics from a theoretical point of view. Our main goal is to analyze them with respect to their runtime behavior. Such analyses can help to understand the working principle of these algorithms on certain problems as well as help to design better algorithms for newly given ones.

Diversity Mechanisms

It is widely assumed and observed in experiments [2, 19] that the use of diversity mechanisms in evolutionary algorithms may have a great impact on its running time. Such mechanisms should ensure that the population of an EA should at each time step contain different search points. We have pointed out the use of different mechanisms with respect to the runtime behavior [12]. The algorithms considered either diversify the population with respect to the search points or with respect to function values. Investigating simple plateau functions, we have shown that using the "right" diversity strategy makes the difference between an exponential and a polynomial runtime.

Combinatorial Optimization

Many problems in combinatorial optimization solvable in polynomial run-time become NP-hard if altered slightly. In application such variants occur frequently. Evolutionary computation methods proved to be good heuristic on several of these variants. Nevertheless, the hardness of these variant problems makes a theoretical analysis difficult. We therefore investigate the behavior of evolutionary search heuristics on the original combinatorial optimization problems.

Evolutionary computing methods have obtained good results on NP-hard variants of the minimum spanning tree (MST) problem. To get an understanding of these heuristics we studied the basic MST problem. We have analyzed simple randomized search heuristics with respect to their runtime behavior. In [16] we have given upper and lower bounds on the run-time different EAs need in expectation to compute a minimum spanning tree.

Another problem we investigated is the cycle cover problem. There are numerous variants of this problem, many of which are NP-hard. We have investigated how to choose the components of an evolutionary search strategy in order to minimize the expected run-time. This

was done by studying the Eulerian cycle problem which can be considered to be the most basic cycle cover problem. First we focused on the choice of the mutation operator. In [15] it is shown, that depending on the mutation operator used, the expected optimization time can be either polynomial or exponential. In [3] and [4], an alternative mutation operator is introduced that improves the expected run-time of the previous result. Two other works shift the focus to the representation of an individual. A new representations of cycles in graphs was presented in [6] and yields a further improvement to the expected run-time of the EAs. We have refined this representation in [5]. The run-time of the resulting EAs differs only by a log-factor from the run-time of specialized algorithms for the Eulerian cycle problem. Since this result is assumed to be best possible this series of run-time studies is exhaustive.

Multi-Objective Optimization

Real-world optimization problems often have many objectives. However, most publications investigate problems where the number of considered objectives is very low. 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.

We have examined how adding objectives to a given optimization problem affects the computation effort required to generate the set of Pareto-optimal solutions [1]. Experimental studies show that additional objectives may change the runtime 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 may be both beneficial and obstructive depending on the chosen objective. Our results are obtained by rigorous runtime analyses that show the different effects of adding objectives to a well-known plateau-function.

We have also considered the approximation ability of randomized search for the class of covering problems and have compared single-objective and multi-objective models for such problems [11]. In contrast to numerous experimental results, there are only a few theoretical results on this subject. For the VertexCover problem, we found situations where the multi-objective model leads to a fast construction of optimal solutions while in the single-objective case even no good approximation can be achieved within expected polynomial time. Examining the more general SetCover problem we show that optimal solutions can be approximated within a factor of $\log n$, where n is the problem dimension using the multi-objective approach while the approximation quality obtainable by the single-objective approach in expected polynomial time may be arbitrarily bad.

Ant Colony Optimization

The theoretical runtime analysis for the modern and very popular randomized search heuristic Ant Colony Optimization (ACO) [9] lags far behind the results for the classical EAs. Until 2006, only convergence results [13], and results on the dynamics of models of ACO [14] were known. In a survey on theoretical studies of ACO [8], researchers were encouraged to follow the approach taken for the analysis of EAs by starting a runtime analysis of simple

ACO algorithms on OneMax. Soon after this appeal, the first theorems on the runtime of a simple ACO algorithm appeared in [18]. In that paper, a simple ACO algorithm called 1-ANT is defined based on the model of [13] and the runtime with respect to the fitness function OneMax is bounded from above and below. It is shown that the so-called evaporation factor ρ, the probably most important parameter in ACO algorithms, has a crucial impact on the runtime.

In [7] we have put forward the analysis of the 1-ANT on non-symmetric example problems in a similar fashion to [18]. We chose the functions LeadingOnes and BinVal investigated in [10] for a well-known EA and analyzed the runtime of the 1-ANT on these functions with respect to n, the dimensionality of the search space, and ρ. It turned out that a similar phase transition behavior can be observed as in [18]. If ρ is asymptotically smaller than a threshold, no efficient optimization is possible; however, for values a little above the threshold, polynomial runtimes are very likely. Our investigations again suggest that the 1-ANT is not robust with respect to the choice of ρ.

We have also presented the first comprehensive rigorous analysis of a simple ACO algorithms for a combinatorial optimization problem [17]. In our investigations we considered the minimum spanning tree problem and examined the effect of two construction graphs with respect to the runtime behavior. The choice of the construction graph in an ACO algorithm seems to be crucial for the success of such an algorithm. First, we took the input graph itself as the construction graph and analyzed the use of a construction procedure that is similar to Broder's algorithm for choosing a spanning tree uniformly at random. After that, a more incremental construction procedure has been analyzed. It turned out that this procedure is superior to the Broder-based algorithm and produces additionally in a constant number of iterations a minimum spanning tree if the influence of the heuristic information is large enough.

References

- [1] 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 (GECCO-2007)*, London, UK, 2007. ACM. To appear.

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

- [3] B. Doerr, N. Hebbinghaus, and F. Neumann. Speeding up evolutionary algorithms through restricted mutation operators. In T. P. Runarsson, H. G. Beyer, E. Burke, J. J. Merelo-Guervós, L. D. Whitley, and X. Yao, eds., *Parallel Problem Solving from Nature - PPSN IX, 9th International Conference*, Reykjavik, Iceland, October 2006, *LNCS 4193*, pp. 978–987. Springer.

- [4] B. Doerr, N. Hebbinghaus, and F. Neumann. Speeding up evolutionary algorithms through unsymmetric mutation operators. *Evolutionary Computation*, 2007. To appear.

- [5] 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. ACM. To appear (nominated for best paper award).

- [6] 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. IEEE. To appear.

- [7] B. Doerr, F. Neumann, D. Sudholt, and C. Witt. On the influence of pheromone updates in aco algorithms. In *Genetic and Evolutionary Computation Conference (GECCO-2007)*, London, UK, 2007. ACM. To appear (nominated for best paper award).

[8] M. Dorigo and C. Blum. Ant colony optimization theory: A survey. *Theor. Comput. Sci.*, 344:243–278, 2005.

[9] M. Dorigo and T. Stützle. *Ant Colony Optimization*. MIT Press, 2004.

[10] S. Droste, T. Jansen, and I. Wegener. On the analysis of the (1+1) evolutionary algorithm. *Theor. Comput. Sci.*, 276:51–81, 2002.

- [11] 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 (GECCO-2007)*, London, UK, 2007. ACM. To appear.

- [12] T. Friedrich, N. Hebbinghaus, and F. Neumann. Rigorous analyses of simple diversity mechanisms. In D. Thierens, ed., *Genetic and Evolutionary Computation Conference (GECCO-2007)*, London, UK, 2007. ACM. To appear (nominated for best paper award).

[13] W. J. Gutjahr. A generalized convergence result for the graph-based ant system metaheuristic. *Probab. Eng. Inform. Sc.*, 17:545–569, 2003.

[14] D. Merkle and M. Middendorf. Modelling the dynamics of Ant Colony Optimization algorithms. *Evolutionary Computation*, 10(3):235–262, 2002.

- [15] F. Neumann. Expected runtimes of evolutionary algorithms for the eulerian cycle problem. *Computers and Operations Research*, 2007. To appear.

- [16] F. Neumann and I. Wegener. Randomized local search, evolutionary algorithms, and the minimum spanning tree problem. *Theoretical Computer Science*, 2007. To appear.

- [17] F. Neumann and C. Witt. Ant colony optimization and the minimum spanning tree problem. Technical Report TR06-143, Electronic Colloquium on Computational Complexity, http://eccc.hpi-web.de/eccc/, November 2006.

[18] F. Neumann and C. Witt. Runtime analysis of a simple ant colony optimization algorithm. In *Proc. of ISAAC '06*, 2006, *LNCS 4288*, pp. 618–627. Springer.

[19] R. K. Ursem. Diversity-guided evolutionary algorithms. In *Proc. of PPSN '02*, 2002, *LNCS 2439*, pp. 462–474.

12.3.3 Computer Algebra

Many algorithmic problems reduce to solving polynomial equations and inequalities. The most fundamental problem of this kind is the solution of one polynomial equation in one variable. We have continued our research on this problem from previous years and obtained new results both in theory and practice.

The Descartes Method for Real Root Isolation

Investigators: Arno Eigenwillig and Michael Kerber, in collaboration with Vikram Sharma and Chee Yap

Given a polynomial $f(x)$ with real coefficients, real root isolation consists in assigning disjoint enclosing intervals to all real roots of $f(x)$. These are called *isolating intervals*. In practice,

one of the best approaches to real root isolation is the Descartes method, whose modern form goes back to Collins and Akritas [1]. It is based on recursive interval bisection combined with Descartes' rule of signs to check whether an interval contains no or exactly one real root. This requires that $f(x)$ has no multiple roots.

In the previous report period, W. Krandick and K. Mehlhorn rediscovered a result that gives an improved termination criterion for the Descartes method. They combined it with the Davenport-Mahler bound on root separation for an improved analysis of the size of the recursion tree. In the meantime, this work has appeared in final form [5]. For a polynomial of degree n with integer coefficients of magnitude less than 2^τ, the bounds attained result in a recursion tree size of $O(n \log n\,(\tau + \log n))$.

A. Eigenwillig, V. Sharma, and C. Yap [4] have improved upon the analysis of Krandick and Mehlhorn. By exploiting geometric properties of the rediscovered termination condition, the Davenport-Mahler bound can be applied in a simpler and more direct way, yielding a bound on tree size of $O(n(\tau + \log n))$. This is almost tight, because a family of explicit, Mignotte-like worst-case inputs necessitates a recursion depth of $\Omega(n\tau)$. This improved bound on recursion tree size entails an improved bit complexity bound of $\tilde{O}(n^4\tau^2)$ for the Descartes method (where $\tilde{O}(\,\cdot\,)$ denotes omission of logarithmic factors).

Also in the previous report period, work had begun on the Bitstream Descartes method, which allows to obtain exact isolating intervals from coefficients that can be approximated as far as needed but are not known exactly. This is achieved by combining a careful analysis of approximation errors with randomization of interval subdivision points to avoid numerically hard cases. During the present report period, this work has appeared in print [3].

M. Kerber makes extensive use of the Bitstream Descartes method in his library AlciX (see §12.8.1) for the geometric analysis of algebraic curves $F(x, y) = 0$. If (α, β) is a critical point of F (e.g., a singularity), β is found as the multiple root of $f(y) = F(\alpha, y)$. In general, the coefficients of $f(y)$ are algebraic irrationals and thus costly in exact computation. The ability to compute with approximate coefficients is a crucial source of efficiency in AlciX. M. Kerber has extended the Descartes method with the necessary treatment of a k-fold root of f, $k > 1$, by a hybrid symbolic-numeric approach. A partial converse of Descartes' rule of the required form (when does the Descartes test count no more than k for a k-fold root?) has been found by A. Eigenwillig [2].

References

[1] G. E. Collins and A. G. Akritas. Polynomial real root isolation using Descartes' rule of signs. In R. D. Jenks, ed., *Proceedings of the 1976 ACM Symposium on Symbolic and Algebraic Computation*, 1976, pp. 272–275. ACM Press.

- [2] A. Eigenwillig. On multiple roots in Descartes' rule and their distance to roots of higher derivatives. *Journal of Computational and Applied Mathematics*, 200(1):226–230, March 2007.

- [3] 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.

- [4] 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, July 2006, pp. 71–78. ACM.

- [5] W. Krandick and K. Mehlhorn. New bounds for the descartes method. *Journal of Symbolic Computation*, 41(1):49–66, 2006.

12.3.4 Game Theory and Mechanism Design

Over the last few years, game theory became an important field for computer scientists. The reason is that nowadays we face more and more systems that are not centrally organized. Examples are the Internet or (electronic) market systems. Often, such system are run by different agents with different objectives. Selfishness or lack of communication leads to these agents trying to maximize their own welfare rather than taking care of the system as whole. Such settings are best modelled in the language of games. The task of setting up rules for such games in a way that the agents behave in the way we want them to is called mechanism design.

Mechanism Design for Scheduling Problems

Investigators: Giorgos Christodoulou and Annamária Kovács in collaboration with Elias Koutsoupias and Angelina Vidali

Scheduling on unrelated machines is a classical NP-hard problem. Lenstra, Shmoys and Tardos [9] 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 [12]. They gave a n-approximation deterministic truthful mechanism and a lower bound of 2. They also conjectured the actual bound to be n.

Scheduling on related machines, from the mechanism design point of view, was first studied by Archer and Tardos [3]. In this variant of the problem, the private parameter for each machine, is a single value (its speed). Archer and Tardos [3] characterized the class of truthful mechanisms for this setting, in terms of a monotonicity condition of the mechanism's allocation algorithm. A similar characterization for one-parameter mechanism design problems (single item auction) can also be found in Myerson [11]. For this problem, it turns out that the optimal allocation algorithm can be modified to be a truthful mechanism. Archer and Tardos [3] gave a randomized truthful 3-approximation algorithm, which was later improved to a 2-approximation by Archer [2]. Andelman, Azar and Sorani [1] gave the first deterministic polynomial mechanism for the problem, with an approximation ratio of 5.

A Lower Bound for Scheduling Mechanisms In [5], we study the mechanism design problem of scheduling tasks on n unrelated machines in which the machines are the players of the mechanism.

The problem is one of the most fundamental scheduling problems [9]. 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. 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 don't 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.

The problem was proposed and studied in the seminal paper of Nisan and Ronen, where it was shown that the approximation ratio of mechanisms is between 2 and n. We improve the lower bound to $1 + \sqrt{2}$ for 3 or more machines.

Fractional Mechanisms for Scheduling on Unrelated Machines In [4], we consider the mechanism design version of fractional scheduling on unrelated machines. 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. Notice 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 [10] to the class of fractional mechanisms. Notice that a fractional mechanism is more powerful than a randomized mechanism for the integral case, since it has the flexibility to split a task into many machines, while a randomized mechanism, finally, has to assign the whole task to a machine, and this affects its approximation ratio.

In the positive direction, we give a truthful mechanism with approximation ratio 3/2 for 2 machines, which matches our lower bound. This is the first new tight bound that we have for any variant of the problem, after the tight bound of 2 in the integral case, obtained for 2 machines in the original paper of Nisan and Ronen [12]. The generalization of our mechanism for n machines gives us an approximation ratio of $1 + \frac{n-1}{2}$.

Next we turn our attention to a family of mechanisms that we call *task-independent*. This family consists of mechanisms, where the decision for the assignment of a task, depends only on the processing times that concern the particular task *(time column that corresponds to the task)*. Considering task-independence is motivated by the fact that all known 'reasonable' deterministic and randomized mechanisms for this problem are task-independent. Furthermore, this sort of independence has attractive properties: easy to design by applying methods for one-parameter auctions, fits well with on-line settings, where tasks may appear one-by-one. It is natural to ask if there is room for improvement on the approximation ratio by use of such mechanisms. We extend this question for the class of task-independent *algorithms* that need not satisfy the additional properties imposed by truthfulness. We give a lower bound of $1 + \frac{n-1}{2}$ on the approximation ratio of any algorithm that belongs to this class. Our mechanism is also task-independent, and hence is optimal over this family of algorithms.

Truthful Mechanisms for Scheduling on Related Machines $Q||C_{\max}$ denotes the problem of scheduling n jobs on m machines of different speeds so that the makespan is minimized.

In [6] we provided a fast and simple, deterministic *monotone* 3-approximation algorithm for $Q||C_{\max}$. Monotonicity is relevant in the context of *truthful mechanisms*: when each machine speed is only known to the machine itself, we need to motivate that machines declare their true speeds to the scheduling mechanism. Such motivation is possible only if the scheduling algorithm used by the mechanism is monotone [11, 3]. The best previous deterministic truthful mechanism (monotone algorithm) that is polynomial in m, was a 5-approximation by Andelman et al [1].

Subsequently [7, 8], we could reduce the approximation ratio of our mechanism to 2.8. We also showed that our mechanism is *frugal*, meaning that the sum of the payments handed to the machines does not exceed the total cost of the agents by more than a logarithmic factor.

References

[1] N. Andelman, Y. Azar, and M. Sorani. Truthful approximation mechanisms for scheduling selfish related machines. In *Proc. 22nd Ann. Symp. on Theor. Asp. of Comp. Sci. (STACS)*, 2005, *LNCS 3404*, pp. 69–82. Springer.

[2] A. Archer. *Mechanisms for Discrete Optimization with Rational Agents*. PhD thesis, Cornell University, January 2004.

[3] A. Archer and E. Tardos. Truthful mechanisms for one-parameter agents. In *Proc. 42nd IEEE Symp. on Found. of Comp. Sci. (FOCS)*, 2001, pp. 482–491.

• [4] G. Christodoulou, E. Koutsoupias, and A. Kovács. Mechanism design for fractional scheduling on unrelated machines. To appear at ICALP 07., 2007.

• [5] 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.

• [6] A. Kovács. Polynomial time preemptive sum-multicoloring on paths. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, eds., *Automata, languages and programming, 32nd International Colloquim, ICALP 2005*, Lisbon, Portugal, July 2005, *LNCS 3580*, pp. 840–852. Springer.

• [7] A. Kovács. Tighter approximation bounds for LPT scheduling in two special cases. In T. Calamoneri, I. Finocchi, and G. F. Italiano, eds., *Algorithms and Complexity, 6th Italian Conference, CIAC 2006*, Rome, Italy, May 2006, *LNCS 3998*, pp. 187–198. Springer.

[8] A. Kovács. *Fast Algorithms for Two Scheduling Problems*. Phd thesis, Universität des Saarlandes, 2007, to appear.

[9] J. Lenstra, D. Shmoys, and É. Tardos. Approximation algorithms for scheduling unrelated parallel machines. *Mathematical Programming*, 46(1):259–271, 1990.

[10] 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.

[11] R. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

[12] 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.

A Quasi-PTAS for Envy-free Pricing on Line Graphs

Investigators: Khaled Elbassioni, René Sitters, and Yan Zhang

We consider the problem of pricing items so as to maximize the profit made from selling these items. An instance is given by a set E of n items and a set of m clients, where each client is specified by one subset of E (the bundle of items he/she wants to buy), and a budget, which is the maximum price he/she wants to pay for the subset. We restrict to the model where the subsets can be arranged such that they form intervals of a line graph. We give a quasi-polynomial time approximation scheme for the capacitated and uncapacitated problems. In the latter problem, also known as the *highway problem*, there is an infinite supply of any item, and hence, we assume that any client who can afford his/her bundle will buy it. In the capacitated problem there is only a limited supply of any item. Hence, we have to set prices as well ad to find a subset of the clients such that the supply constraints are satisfied. This variant of the problem is closely related to the unsplittable flow problem on line graphs for which a quasi-PTAS was developed only recently.

Mechanisms for Product Pricing

Investigators: René Sitters in collaboration with Alexander Grigoriev, Joyce van Loon, and Marc Uetz

We study algorithms for the multi-product pricing problem, where, given consumer preferences among products, their budgets, and the costs of production, the goal is to set prices of multiple products from a single company, so as to maximize the overall revenue of the company. We present approximation algorithms as well as negative results for several variants of the multi-product pricing. In our model, each potential buyer j is interested in one specific subset of the items and is willing to pay a maximum price b_j for this set [1]. A client j will buy his preferred set for a price x if and only if the prices add up to $x \leq b_j$. In the exceptional event that all sets are disjunct, then the seller can earn the maximum price b_j from each client j. Naturally, clients have similar interests which makes the profit maximization problem much harder to solve. The items are modeled as edges of a graph and each budget forms a simple path. If the underlying graph is a path and there is only a fixed number of each item available, we derive a fully polynomial time approximation scheme, complementing a recent NP-hardness result. If the graph is a tree we show that the problem is polynomially solvable, contrasting its APX-hardness for the case of unlimited availability of items. On the negative side, we show that it is NP-hard to approximate the problem within a factor $n^{1?\epsilon}$ even if the graph takes the form of a grid.

References

- [1] A. Grigoriev, J. v. Loon, R. Sitters, and M. Uetz. How to sell a graph: Guidelines for graph retailers. In F. V. Fomin, ed., *32nd International Workshop on Graph-Theoretical Concepts in Computer Science, WG 2006*, Bergen, Norway, October 2006, *LNCS 4271*, pp. 125–136. Springer.

12.3.5 Graph Theory and Algorithms

Graphs are a fundamental mathematical abstraction of the real-world phenomenon of being neighbors. They allow beautiful theory like Euler's famous solution to the Königsberg Bridge problem [1] and are the basis of many classical algorithmic problems like finding shortest paths.

References

[1] L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741.

Minimum Cycle Bases

Viewing graphs in an algebraic fashion, we can define several vector spaces on a graph. Since edges carry most of the structure of a graph, we are mostly concerned with the edge space, the vector space formed by the edge sets of a graph. One of the most important subspaces of the edge space is the cycle space.

For undirected graphs, a $\{0,1\}$ incidence vector is associated with each cycle and the vector space over $\mathbb{F}_2$ generated by these vectors is the cycle space of G. A set of cycles is called a cycle basis of G if it forms a basis for its cycle space. A cycle basis where the sum of the weights of the cycles is minimum is called a minimum cycle basis of G (abbreviated as MCB). Similarly, for directed graphs, cycles are associated with $\{-1,0,1\}$ incidence vectors: edges traversed by the cycle in the right direction get 1, edges traversed in the opposite direction get -1, and edges not in the cycle at all get 0. The vector space generated by these incidence vectors over $\mathbb{Q}$ is the cycle space. The books by Deo [3] and Bollobás [2] contain an in-depth coverage of cycle bases.

One of the most important areas of application of the MCB problem is electric networks [9, 8, 1]. Many problems arising in the design and analysis of electric networks can be formulated in graph-theoretic terms. In fact, in the analysis of complex electric networks by graph theoretical methods, a basic problem is to determine the solvability of the "network equation", a system of algebraic differential equations that describes the relation of currents and voltages in a network as functions of time. In order to check structural solvability of that system quickly by a heuristic matching approach, fast algorithms are needed that compute sparse representations. The equations corresponding to the Kirchhoff voltage law are critical, since for the remaining equations, a sparse representation is readily available. Hence the central problem is that of computing a sparse cycle basis to describe the "voltage law" part of the system.

Cycle bases of low weight are also useful in a number of other contexts, e.g. structural engineering, (bio)chemistry, and surface reconstruction.

Exact Algorithms

Investigators: Kurt Mehlhorn in collaboration with Ramesh Hariharan and Telikepalli Kavitha

There are several algorithms for computing a minimum cycle basis in an undirected graph and the fastest runs in $O(m^2n+mn^2\log n)$ time [6]. The framework used in these algorithms was introduced by de Pina [8] and was also used in [1, 6]: we compute cycles C_i and supporting vectors N_i so that each C_i is a shortest cycle *not* orthogonal to its corresponding N_i, and each N_i is orthogonal to all previous C_j, $j < i$. This collection of cycles C_i is known to be a minimum cycle basis.

Using this framework we presented the first polynomial time algorithm for computing a minimum cycle basis in a directed graph with running time $\tilde{O}(m^4n)$ [10]. Liebchen and Rizzi [7] gave an $\tilde{O}(m^{\omega+1}n)$ algorithm for this problem, where $\omega < 2.376$ is the exponent of matrix multiplication; this was the fastest deterministic algorithm so far for this problem in directed graphs.

In [4] we present an $O(m^3n + m^2n^2\log n)$ deterministic algorithm and an $O(m^2n + mn^2\log n)$ Monte Carlo algorithm to compute a minimum cycle basis in a directed graph G with m edges, n vertices and non-negative edge weights. The running time of our deterministic algorithm is m times the running time of the fastest algorithm for computing minimum cycle bases in undirected graphs and we leave it as a challenge to close the gap. The increased complexity seems to stem from the larger base field. Arithmetic in $\mathbb{F}_2$ suffices for undirected graphs. For directed graphs, the base field is $\mathbb{Q}$ and this seems to necessitate the handling of large numbers. Also, the computation of a shortest cycle that has a non-zero dot product with a given vector seems more difficult in directed graphs than in undirected graphs.

Approximation Algorithms

Investigators: Kurt Mehlhorn and Dimitrios Michail in collaboration with Telikepalli Kavitha

Many applications which use cycle bases, do not necessarily require a minimum basis but just a sparse one. The running time of current MCB algorithms is still quite high and, furthermore, these algorithms use fast matrix multiplication. For these reasons we investigated the problem of computing approximate minimum cycle bases.

First, we found a number of approximation algorithms which find near minimal bases [5]. The approximations are within a constant factor of optimum. Some of them also work for directed graphs. We present a new approximation approach which leads to vastly improved time bounds. In particular, for any integer $k \geq 1$, we give two $(2k-1)$ approximation algorithms with expected running time $O(kmn^{1+2/k} + mn^{(1+1/k)(\omega-1)})$ and deterministic running time $O(n^{3+2/k})$, respectively. Here ω is the best exponent of matrix multiplication. It is presently known that $\omega < 2.376$. Both algorithms are $o(m^\omega)$ for sufficiently dense graphs, the first algorithm for number of edges $m > \max(n^{1+1/k}, \sqrt[\omega-1]{kn^{1+2/k}})$ and the second algorithm for $m > n^{\frac{3}{\omega}+\frac{2}{k\omega}} = n^{1.26+\frac{0.84}{k}}$. The first algorithm is faster for sparser graphs and the second algorithm for denser graphs. More precisely, the second algorithm is faster for $m > n^{4-\omega+\frac{3-\omega}{k}}$ which with the current upper bound on ω is $m > n^{1.624+\frac{0.624}{k}}$. For

directed graphs we give a $(2k-1)$-approximation deterministic algorithm with an $O(n^{4+3/k})$ running time and a randomized one with $O(n^{3+2/k})$.

Our algorithms work in two phases. The first phase is a very fast computation of a large number of cycles (all but $O(n^{1+1/k})$ cycles) in an approximate MCB. The second part is a more expensive computation of the remaining cycles. We present two different ways for computing these remaining cycles, leading to the above two algorithms, each faster for different graph densities. Only the second phase needs a null space computation; it is a null space computation of a square system of size $O(n^{1+1/k})$. Our new algorithms are fast even when implemented without fast matrix multiplication. Furthermore, by combining the techniques of both the algorithms, we get an even faster algorithm at the expense of a larger approximation factor.

We also present a 2-approximation algorithm with $O(m^\omega \sqrt{n \log n})$ expected running time. For sparse graphs, this is subcubic. Moreover, we develop very fast approximation algorithms for some special graph classes. For planar graphs we give a linear time 2-approximation algorithm and for the complete Euclidean graph in the plane we give a 2.42-approximation algorithm with running time $O(n^3)$. In higher dimensions we give a k-approximation algorithm for any $k > 1$ with running time $O(s^d n^3 \log n)$ where $s = 4(k+1)/(k-1)$ and d is the fixed dimension.

References

[1] F. Berger, P. Gritzmann, and S. de Vries. Minimum cycle basis for network graphs. *Algorithmica*, 40(1):51–62, 2004.

[2] B. Bollobas. *Modern Graph Theory*. Springer-Verlag, 1998.

[3] N. Deo. *Graph Theory with Applications to Engineering and Computer Science*. Prentice Hall, 1974.

• [4] R. Hariharan, K. Telikepalli, and K. Mehlhorn. A faster deterministic algorithm for minimum cycle bases in directed graphs. 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. 250–261. Springer.

• [5] T. Kavitha, K. Mehlhorn, and D. Michail. New approximation algorithms for minimum cycle bases of graphs. In W. Thomas and P. Weil, eds., *24th Annual Symposium on Theoretical Aspects of Computer Science (STACS 2007)*, Aachen, Germany, February 2007, *LNCS 4393*, pp. 512–523. Springer.

[6] T. Kavitha, K. Mehlhorn, D. Michail, and K. Paluch. A faster algorithm for minimum cycle basis of graphs. In J. Díaz, J. Karhumäki, A. Lepistö, and D. Sannella, eds., *Automata, languages and programming, 31st International Colloquium, ICALP 2004*, Turku, Finland, July 2004, *LNCS 3142*, pp. 846–857. Springer.

[7] C. Liebchen and R. Rizzi. A greedy approach to compute a minimum cycle basis of a directed graph. *Inf. Process. Lett.*, 94(3):107–112, 2005.

[8] J. de Pina. *Applications of Shortest Path Methods*. PhD thesis, University of Amsterdam, Netherlands, 1995.

[9] M. N. S. Swamy and K. Thulasiraman. *Graphs, Networks, and Algorithms*. John Wiley & Sons, New York, 1981.

• [10] K. Telikepalli and K. Mehlhorn. A polynomial time algorithm for minimum cycle basis in directed graphs. In V. Diekert and B. Durand, eds., *STACS 2005, 22nd Annual Symposium on Theoretical Aspects of Computer Science*, Stuttgart, Germany, 2005, *LNCS 3404*, pp. 654–665. Springer.

Sensitivity Analysis of Minimum Spanning Trees and Shortest Path Trees

Investigator: Seth Pettie

Every optimization problem has a twin problem concerning the *sensitivity* of its solutions under perturbations to the input weights. For instance, an algorithm computing the sensitivity of a minimum spanning tree returns the amount by which each individual edge weight could be increased or decreased without affecting the identity of the minimum spanning tree. Tarjan [4] used the *path compression* technique to find the sensitivity of minimum spanning trees and single source shortest path trees in $O(m\alpha(m,n))$ time, where m is the number of edges, n the number of vertices, and α the inverse-Ackermann function.

In [2] we designed a faster and substantially simpler algorithm for computing the sensitivity of minimum spanning trees and shortest path trees. The algorithms run in $O(m \log \alpha(m,n))$ time and are based on an improved *split-findmin* data structure. This improvement has also led to asymptotically faster algorithms for computing all-pairs shortest paths in directed and undirected graphs [3, 1].

References

[1] S. Pettie. On the comparison-addition complexity of all-pairs shortest paths. In *Proc. 13th Int'l Symp. on Algorithms and Computation (ISAAC)*, 2002, pp. 32–43.

• [2] S. Pettie. Sensitivity analysis of minimum spanning trees in sub-inverse-Ackermann time. In *Proceedings 16th Int'l Symposium on Algorithms and Computation (ISAAC)*, 2005, pp. 964–973.

• [3] S. Pettie and V. Ramachandran. A shortest path algorithm for real-weighted undirected graphs. *SIAM J. Comput.*, 34(6):1398–1431, 2005.

[4] R. E. Tarjan. Sensitivity analysis of minimum spanning trees and shortest path problems. *Inf. Process. Lett.*, 14(1):30–33, 1982. See Corrigendum, IPL **23**(4):219.

Generalized Minimum Spanning Tree Problems

Investigators: René Sitters in collaboration with Hans Bodlaender, Corinne Feremans, Alexander Grigoriev, Eelko Penninkx, and Thomas Wolle

Connecting a set of points in the plane by a network of minimum length is an easy problem. It becomes much harder to solve if we want to connect subsets of the plane instead of single points. In two papers we investigate the complexity of generalized minimum spanning tree problems. In [1] we consider the *minimum corridor connection problem* where for a given decomposition of a polygon into "rooms", one has to find the minimum length tree along the edges of the decomposition such that every room is incident to a vertex of the tree. We show that the problem is strongly NP-hard and give a sub-exponential time exact algorithm. We develop a polynomial time approximation scheme for the case when all rooms are fat and have nearly the same size. When rooms are fat but are of varying size we give a polynomial

time constant factor approximation algorithm. In [2], we search for the boundary between the easy problem of spanning single points in the plane, and the hard problem of spanning subsets. We show that finding the minimum spanning tree is NP-hard even if each subset contains at most two points and, moreover, these points are close in the following way: Every pair is contained in one cell of an underlying unit grid and, vice versa, the points within one cell form a subset. Alternatively, NP-hardness holds if all sets are pairs with the following local property: for any point, its nearest neighbor is the other point from the pair.

References

- [1] H. L. Bodlaender, C. Feremans, A. Grigoriev, E. Penninkx, R. Sitters, and T. Wolle. On the minimum corridor connection and other generalized geometric problems. In T. Erlebach and C. Kaklamanis, eds., *4th Workshop on Approximation and Online Algorithms, WAOA*, Zurich, Switzerland, September 2006, *LNCS 4368*, pp. 69–82. Springer.
- [2] C. Feremans, A. Grigoriev, and R. Sitters. The geometric generalized minimum spanning tree problem with grid clustering. *4OR: A Quarterly Journal of Operations Research*, 4(4):319–329, 2006.

Graph Isomorphism

Investigators: Martin Kutz and Pascal Schweitzer

A graph isomorphism is a bijection from a vertex set of one graph to the vertex set of a second graph which maps pairs of adjacent vertices to pairs of adjacent vertices and pairs of non-adjacent vertices to pairs of non-adjacent vertices. The Graph Isomorphism problem decides whether there exists a graph isomorphism for a given pair of graphs. Essentially all other isomorphism questions of combinatorial structures can be reduced to Graph Isomorphism. The problem is neither known to be in P nor known to be NP-complete. In [1] we present a Monte Carlo algorithm that can determine with high probability that two graphs are not isomorphic. Our algorithm randomly samples substructures in the given graphs in order to detect dissimilarities between them. On difficult instances it runs faster than state of the art isomorphism solvers, however we do not supply any bound for the running time. The answer is certified in the sense that given an isomorphic pair of graphs an isomorphism is supplied, otherwise a certificate is provided that can be verified in a randomized fashion, with time consumption considerably less than for the search process for the certificate.

References

- [1] M. Kutz and P. Schweitzer. ScrewBox: a randomized certifying graph non-isomorphism algorithm. In D. Applegate and G. Brodal, eds., *9th Workshop on Algorithm Enginneering and Experiments (ALENEX'07)*, New Orleans, USA, 2007. SIAM.

Sampling Random Planar Graphs

Investigator: Daniel Johannsen

Average case run-time analysis is a central research topic in complexity theory. For problems in combinatorial optimization it is closely related to the study of random graphs. Under-

standing the structure of a random graph is essential to determine the behavior of a graph algorithm on a random instance.

Planarity is an elementary property in graph theory. Planar graphs model real word structures, for example single layered computer chips. While the structural properties of planar graphs have been well studied, it is not obvious how to generate a planar graph uniformly at random.

In [1, 2] an decomposition strategy for 3-connected planar graphs is presented. This decomposition leads to a deterministic and polynomial time algorithm to sample labeled planar graphs uniformly at random. The decomposition scheme also allows for the enumeration of labeled 3-connected planar graphs, both recursively and in terms of generating functions. The ongoing research focuses on extending these results to unlabeled planar graphs.

References

- [1] D. Johannsen. Sampling rooted 3-connected planar graphs in deterministic polynomial time. Masters thesis, Humboldt-Universität zu Berlin, April 2006.
- [2] D. Johannsen. A direct decomposition of 3-connected planar graphs. *Séminaire Lotharingien de Combinatoire*, 54A:15 pp, 2007.

Online Topological Ordering

Investigators: Deepak Ajwani, Tobias Friedrich, and Ulrich Meyer

Online topological ordering and online cycle detection are important steps for many applications like pointer analysis and incremental compilation. The problem was first introduced by Alpern et al. [3] in the context of incremental evaluation of computational circuits. For inserting a sequence of m edges, Marchetti Spaccamela et al. [6] and Katriel and Bodlaender [4, 5] proposed different algorithms and analyzed their worst-case runtimes. Pearce and Kelly [7] gave an algorithm that empirically outperforms all others on sparse random DAGs.

We have given a simple algorithm which maintains the topological order of a directed acyclic graph with n nodes under an online edge insertion sequence in $O(n^{2.75})$ time, independent of the number of edges m inserted [2]. For dense DAGs, this is an improvement over the previous best result of $O(\min\{m^{\frac{3}{2}} \log n, m^{\frac{3}{2}} + n^2 \log n\})$ by Katriel and Bodlaender [4, 5]. Empirically, our implementation outperforms the other algorithms on certain hard instances while it is still competitive on random edge insertion sequences leading to complete DAGs.

All known algorithms for online topological ordering are either only analyzed for worst-case insertion sequences or only evaluated experimentally on random DAGs. In [1] we have done the first average-case analysis of online topological ordering algorithms and proved an expected runtime of $\tilde{O}(n^2)$ under insertion of the edges of a complete DAG in a random order for the algorithms of Alpern et al. [3], Katriel and Bodlaender [5], and Pearce and Kelly [7]. This is much less than the best known worst-case bound $O(n^{2.75})$ for this problem.

References

- [1] D. Ajwani and T. Friedrich. Online topological ordering under random edge insertions. Submitted, 2007.

- [2] D. Ajwani, T. Friedrich, and U. Meyer. An o($n^{2.75}$) algorithm for online topological ordering. In L. Arge and R. Freivalds, eds., *Algorithm theory - SWAT 2006, 10th Scandinavian Workshop on Algorithm Theory*, Riga, Latvia, 2006, *LNCS 4059*, pp. 53–64. Springer.

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

- [4] I. Katriel and H. L. Bodlaender. Online topological ordering. In *Proceedings of the sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-05)*, Vancouver, Canada, 2005, pp. 443–450. SIAM.

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

[6] A. Marchetti-Spaccamela, U. Nanni, and H. Rohnert. Maintaining a topological order under edge insertions. *Information Processing Letters*, 59(1):53–58, 1996.

[7] D. J. Pearce and P. H. J. Kelly. A dynamic algorithm for topologically sorting directed acyclic graphs. In *Proceedings of the Workshop on Efficient and experimental Algorithms*, 2004, *LNCS 3059*, pp. 383–398. Springer-Verlag.

Low Outdegree Orientations of Undirected Graphs with Applications

Investigator: Lukasz Kowalik

This work concerns the problem of finding such an orientation of a given undirected graph that the largest number of edges leaving a vertex (called the outdegree of the orientation) is small.

For any $\varepsilon \in (0,1)$ we show [12] an $\tilde{O}(|E(G)|/\varepsilon)$ time algorithm[1] which finds an orientation of an input graph G with outdegree at most $\lceil(1+\varepsilon)d^*\rceil$, where d^* is the maximum density of a subgraph of G. It is known that the optimal value of orientation outdegree is $\lceil d^*\rceil$.

Our algorithm is based on network flows. The key observation is that the classical Dinitz's network flow algorithm used to check whether a given graph admits k-orientation works very fast if k is far enough from optimum value. Thus approximation guarantee given on the input of the algorithm gives an upper bound on running time of Dinitz's algorithm. We use binary search to find such k that Dinitz's algorithm finds k-orientation fast enough.

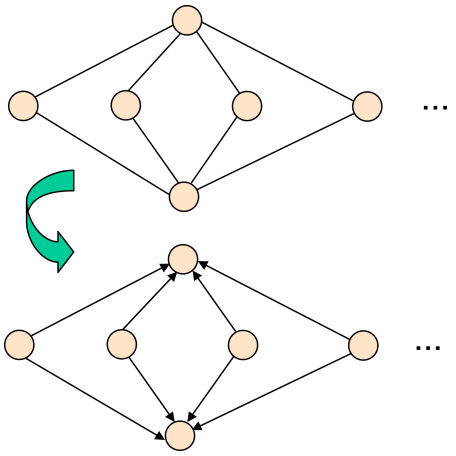


Figure 12.2: Orientation of $K_{2,n}$ with outdegree 2

Our algorithm has applications in constructing labeling schemes, introduced by Kannan *et al.* in [11]. More precisely, for a given graph one can assign labels to vertices so that adjacency of two vertices can be inferred using only their

[1] The $\tilde{O}(\cdot)$ notation ignores logarithmic factors.

labels. In particular for sparse graphs, like e.g. planar graphs, it is possible to find labels of $O(\log |V(G)|)$ bits. (Note that to store a number of a vertex one needs $\lceil |V(G)| \rceil$ bits). Using our algorithm one can find labels of $\lceil (1+\varepsilon)d^* \rceil \cdot \lceil \log |V(G)| \rceil$ bits, in time $\tilde{O}(|E(G)|/\varepsilon)$.

We apply our algorithm also to approximation of such graph density measures as arboricity, pseudoarboricity and maximum density (i.e. density of the densest subgraph). We obtain approximation schemes with additional additive errors of 2, 1 and 1 respectively. The time complexity of these algorithms matches the time complexity of the main algorithm. In this sense, our results improve over the previous, linear-time 2-approximation algorithms by Aichholzer *et al.* [1] (for orientation / pseudoarboricity), by Arikati *et al.* [2] (for arboricity) and by Charikar [5] (for maximum density).

Edge-Coloring

Investigators: Lukasz Kowalik in collaboration with Richard Cole and Riste Škrekovski

In the problem of edge-coloring the input is an undirected graph and the task is to assign colors to the edges so that edges with a common endpoint have different colors. This is one of the most natural graph coloring problems and arises in a variety of scheduling applications.

We consider the problem of verifying whether a given graph G is edge-colorable with k colors and finding such a coloring. Let $\Delta(G)$ denote the maximum degree in graph G. Trivially at least $\Delta(G)$ colors are needed, so if $k < \Delta(G)$ the answer is "no". On the other hand, Vizing [14] proved that when $k \geq \Delta + 1$ the answer is "yes". Unfortunately, when $k = \Delta(G)$ the problem is NP-hard even for $k \geq 3$.

Recently there is a growing interest in finding nontrivial exponential algorithms for NP-hard problem and the goal is to minimize the exponent in the complexity function. We follow this line of research and we focus on the simplest NP-hard case: $k = 3$.

One way to solve our problem is to apply a vertex-coloring algorithm to the line graph L of G, i.e. the graph with vertices corresponding to the edges of G and with edges describing the incidence relation. The currently fastest algorithm for 3-vertex-coloring a given graph L is due to Beigel and Eppstein [3] and it works in $O(1.3289^{|V(L)|})$ time. For $k = 3$, since $\Delta(G) = 3$, the line graph has at most $\frac{3}{2}n$ vertices (n denotes the number of vertices in the input graph G), hence it yields an $O(1.532^n)$ algorithm. However, for 3-edge-coloring Beigel and Eppstein get an $O(2^{n/2}) = O(1.415^n)$-time algorithm.

We presented [13] a 3-edge-coloring algorithm with time complexity $O(1.344^n)$ $= O(2^{0.427n})$. The space complexity of our algorithm is polynomial (even linear). We apply the "measure and conquer" technique (developed in [3] and [10]), which suggests to design a relatively simple algorithm and then provide a nontrivial time complexity analysis by introducing a non-standard instance size function. Our algorithm extends a very natural approach of generating inclusion-maximal matchings of the graph.

For planar graphs the problem of edge-coloring is no longer NP-hard for instances with the maximum degree Δ large enough. More precisely, it is known that for $\Delta \geq 7$ it is possible to edge-color planar graph with Δ colors and the famous conjecture of Vizing states that it is still true for $\Delta = 6$ (while we have examples that $\Delta + 1$ is sometimes needed for smaller values of Δ). We focused on algorithms for finding such colorings. We designed a linear-time algorithm [8] for coloring planar graphs with maximum degree Δ with $\max\{\Delta, 9\}$ colors. This improve over the algorithms of Chrobak and Yung [7] and of Chrobak and Nishizeki [6]

which color planar graphs using $\max\{\Delta, 19\}$ colors in linear time or using $\max\{\Delta, 9\}$ colors in $\mathcal{O}(n \log n)$ time. Moreover for $\Delta = 4, 5, 6$ we described linear-time algorithms that use $\Delta + 2$ colors.

We also considered another natural variant of the problem where the color used to color an edge has to be chosen from a list of allowed colors (each edge has assigned its own list). This is so-called list-edge-coloring. It was already known [4] that for $\Delta \geq 12$ any choice of lists of length at least Δ admits such coloring. In paper [9] we described a linear time algorithm which finds it. We obtained this algorithm as a by-product of main result of the paper which is a generalization of Kotzig's Theorem which says that any planar graph with minimum degree 3 contains an edge with sum of its ends degrees bounded by 13. We showed that when the degree constraint is decreased to 2 and such an edge does not exist then the graph contains one of 5 small subgraphs of regular structure, called crowns.

References

[1] O. Aichholzer, F. Aurenhammer, and G. Rote. Optimal graph orientation with storage applications. SFB-Report F003-51, SFB 'Optimierung und Kontrolle', TU Graz, Austria, 1995.

[2] S. R. Arikati, A. Maheshwari, and C. D. Zaroliagis. Efficient computation of implicit representations of sparse graphs. *Discrete Appl. Math.*, 78(1-3):1–16, 1997.

[3] R. Beigel and D. Eppstein. 3-coloring in time $O(1.3289^n)$. *J. Algorithms*, 54(2):168–204, February 2005.

[4] O. V. Borodin, A. V. Kostochka, and D. R. Woodall. List edge and list total colourings of multigraphs. *Journal of Combinatorial Theory, Series B*, 71:184–204, 1997.

[5] M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In *Proc. 13th Int. Workshop on Approximation Algorithms for Combinatorial Optimization (APPROX'00)*, 2000, *LNCS 1913*, pp. 84–95.

[6] M. Chrobak and T. Nishizeki. Improved edge-coloring algorithms for planar graphs. *Journal of Algorithms*, 11:102–116, 1990.

[7] M. Chrobak and M. Yung. Fast algorithms for edge-coloring planar graphs. *Journal of Algorithms*, 10:35–51, 1989.

• [8] R. Cole and L. Kowalik. New linear-time algorithms for edge-coloring planar graphs. *Algorithmica*, 2007. To appear.

• [9] R. Cole, L. Kowalik, and R. Skrekovski. A generalization of Kotzig's theorem and its application. *SIAM Journal on Discrete Mathematics*, 21:93–106, 2007.

[10] F. Fomin, F. Grandoni, and D. Kratsch. Measure and conquer: Domination – a case study. In *Proc. 32nd International Colloquium on Automata, Languages and Programming (ICALP'05)*, 2005, pp. 191–203.

[11] S. Kannan, M. Naor, and S. Rudich. Implicit representation of graphs. In *Proc. of the 20th Annual ACM Symposium on Theory of Computing (STOC '88)*, New York, NY, USA, 1988, pp. 334–343. ACM Press.

• [12] L. Kowalik. Approximation scheme for lowest outdegree orientation and graph density measures. In T. Asano, ed., *Algorithms and Computation, 17th International Symposium, ISAAC 2006*, Kolkata, India, 2006, *LNCS 4288*, pp. 557–566. Springer.

• [13] L. Kowalik. Improved edge coloring with three colors. In F. V. Fomin, ed., *Graph-Theoretic Concepts in Computer Science, 32nd International Workshop, WG 2006*, Bergen, Norway, 2006, *LNCS 4271*, pp. 90–101. Springer.

[14] V. G. Vizing. On the estimate of the chromatic class of a p-graph. *Diskret. Analiz*, 3:25–30, 1964.

12.3.6 Data Structures

Data structures are the backbone of algorithms. Sorting and string problems are some of the very basic problems. Although the basic questions have been answered, there are still many open problems in these areas which remain relevant in our days.

How Branch Mispredictions Affect Quicksort

Investigator: Kanela Kaligosi in collaboration with Peter Sanders

Sorting is one of the most important algorithmic problems both practically and theoretically. Quicksort [2] is perhaps the most frequently used sorting algorithm since it is very fast in practice, needs almost no additional memory, and makes no assumptions on the distribution of the input. Hence, quicksort, its analysis and efficient implementation is discussed in most basic courses on algorithms. When we take a random pivot, the expected number of comparisons is $2n \ln n \approx 1.4n \lg n$. One of the most well known optimizations is that taking the median of three elements reduces the expected number of comparisons to $\frac{12}{7} n \ln n \approx 1.2n \lg n$ [4]. Indeed, by using the median of a larger random sample, the expected number of comparisons can be made as close to $n \lg n$ as we want [5]. At first glance, counting comparisons makes a lot of practical sense since in quicksort, the number of executed instructions and cache faults grow proportionally with this figure.

However, in comparison based sorting algorithms like quicksort or mergesort, neither the executed instructions nor the cache faults dominate execution time. Comparisons are much more important, but only indirectly since they cause the direction of branch instructions depending on them to be mispredicted. In modern processors with long execution pipelines and superscalar execution, dozens of subsequent instructions are executed in parallel to achieve a high peak throughput. When a branch is mispredicted, much of the work already done on the instructions following the predicted branch direction turns out to be wasted. Therefore, ingenious and very successful schemes have been devised to accurately *predict* the direction a branch takes.

Our main theoretical contribution is an analysis of quicksort in the context of branch mispredictions. For simplicity we assume that the elements are distinct. We look at two variants of quicksort: random and skewed pivot, and three branch prediction methods: static, 1-bit predictor and 2-bit predictor. To the best of our knowledge this represents the first analysis of the interactions of a nontrivial algorithm with dynamic branch prediction methods. The theoretical results are complemented by experiments. In particular, we also look at the classical median-of-three pivot selection. It turns out that this frequently used improvement only gives a negligible advantage over random pivot. Its advantages wrt. instruction count basically cancel with its disadvantages wrt. branch prediction. Somewhat surprisingly, taking a pivot with rank around $n/10$ can lead to a better performance.

Figs. 12.3, 12.4 and 12.5 show a comparison of the random pivot, the median of 3, the exact median or 1/2-skewed and the 1/10-skewed pivoting mechanisms in terms of the execution time, the number of occurring branch mispredictions and the number of instructions executed for different values of n. For comments on the figures we refer the reader to [3].

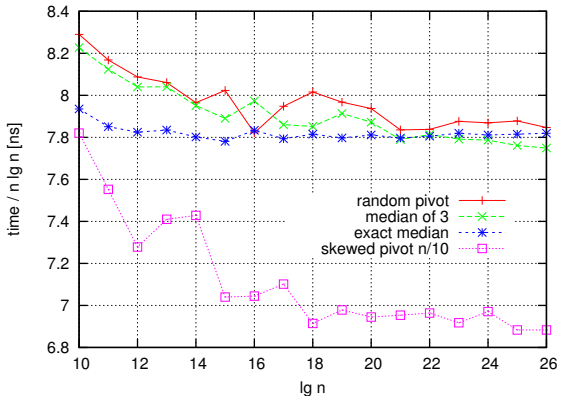


Figure 12.3: Time / $n \lg n$ for random pivot, median of 3, exact median, 1/10-skewed pivot

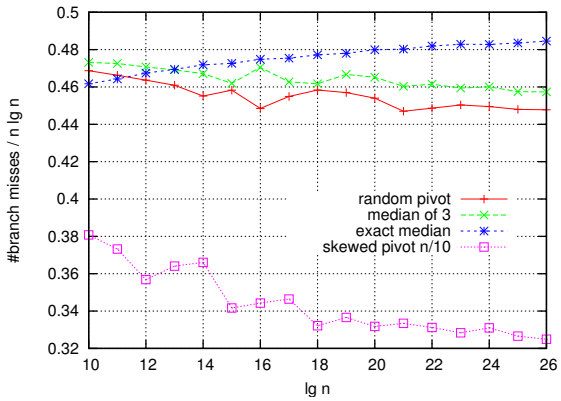


Figure 12.4: Number of branch mispredictions / $n \lg n$ for random pivot, median of 3, exact median, 1/10-skewed pivot

Faster Algorithms for Computing Longest Common Increasing Subsequences

Investigators: Kanela Kaligosi and Martin Kutz in collaboration with Gerth Stoelting Brodal and Irit Katriel

Algorithms that search for the longest common subsequence (LCS) of two input sequences or the longest increasing subsequence (LIS) of one input sequence date back several decades.

Formally, given two sequences $A = (a_1, \ldots, a_n)$ and $B = (b_1, \ldots, b_m)$ with elements from an alphabet Σ and with $m \geq n$, a *common subsequence* of A and B is a subsequence $(a_{j_1} = b_{\kappa_1}, a_{j_2} = b_{\kappa_2}, \ldots a_{j_\ell} = b_{\kappa_\ell})$, where $j_1 < j_2 < \cdots < j_\ell$ and $\kappa_1 < \kappa_2 < \cdots < \kappa_\ell$.

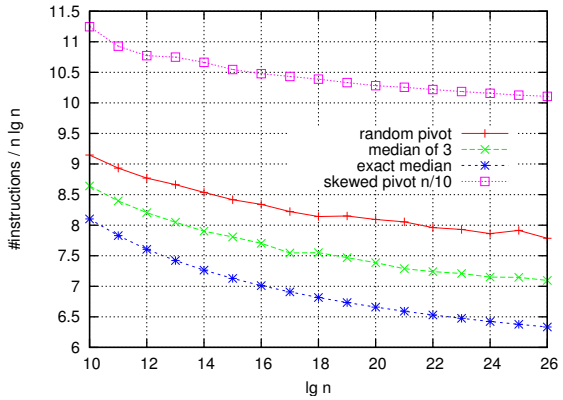


Figure 12.5: Number of instructions / $n \lg n$ for random pivot, median of 3, exact median, 1/10-skewed pivot

Given one sequence $A = (a_1, \ldots, a_n)$ where the a_i's are drawn from a totally ordered set, an *increasing subsequence* of A is a subsequence $(a_{j_1}, a_{j_2}, \ldots, a_{j_\ell})$ such that $j_1 < j_2 < \cdots < j_\ell$ and $a_{j_1} < a_{j_2} < \cdots < a_{j_\ell}$.

Recently, Yang et al. [6] combined the two concepts and defined a *common increasing subsequence* (CIS) of two sequences A and B, i.e., an increasing sequence that is a subsequence of both A and B. They designed a dynamic programming algorithm that finds a *longest CIS* (an LCIS, for short) of A and B using $\Theta(mn)$ time and space.

In [1] we present three new upper bounds for the LCIS problem. The first is an output-dependent algorithm which runs in $O((m+n\ell)\log\log\sigma + Sort_\Sigma(m))$ expected time and $O(m)$ worst-case space, where ℓ is the length of an LCIS. Whenever $n = \Omega(\log\log\sigma + Sort_\Sigma(m)/m)$ and either $m = \Omega(n\log\log\sigma)$ or $\ell = o(n/\log\log n)$, it is faster than Yang et al.'s $\Theta(mn)$-time algorithm.

For a strictly-increasing subsequence we have $\ell \leq \sigma$. However, in the weakly-increasing (i.e. non-decreasing) variant, the length of the output can be arbitrarily larger than the size of the alphabet. We show that a *longest common weakly increasing subsequence* (LCWIS) can be found in linear time for an alphabet of size two and in $O(m + n\log n)$ time for an alphabet of size three. These results are interesting because they pinpoint what seems to be a fundamental difference between the LCS and LWCIS problems. The approach we use cannot be applied to LCS, and to date, comparable speedups have not been achieved for LCS with small alphabets.

Finally, we consider the case of $k \geq 3$ length-n sequences. The upper bound of Chan et al. is achieved by two algorithms; the first is a simple $O(kr^2 + kSort_\Sigma(n))$ time algorithm and the second is a more complex implementation of the same approach, which runs in $O(kr\log\sigma\log^{k-1} r + kSort_\Sigma(n))$ time. We describe an algorithm which is significantly simpler than the latter and obtain a running time of $O(\min\{kr^2, r\log^{k-1} r\log\log r\} + kSort_\Sigma(n))$.

References

- [1] G. S. Brodal, K. Kaligosi, I. Katriel, and M. Kutz. Faster algorithms for computing longest common increasing subsequences. In L. Moshe and V. Gabriel, eds., *Combinatorial Pattern*

Matching, 17th Annual Symposium, CPM 2006, Barcelona, Spain, 2006, *LNCS 4009*, pp. 330–341. Springer.

[2] C. A. R. Hoare. Algorithm 64: Quicksort. *Commun. ACM*, 4(7):321, 1961.

• [3] K. Kaligosi and P. Sanders. How branch mispredictions affect quicksort. In Y. Azar and T. Erlebach, eds., *Algorithms - ESA 2006, 14th Annual European Symposium*, Zürich, Switzerland, 2006, *LNCS 4168*, pp. 780–791. Springer.

[4] D. E. Knuth. *The Art of Computer Programming—Sorting and Searching*, vol. 3. Addison Wesley, 1973.

[5] C. Martínez and S. Roura. Optimal sampling strategies in Quicksort and Quickselect. *SIAM Journal on Computing*, 31(3):683–705, June 2002.

[6] I.-H. Yang, C.-P. Huang, and K.-M. Chao. A fast algorithm for computing a longest common increasing subsequence. *Inf. Process. Lett.*, 93/5:249–253, 2005.

Pairing Heaps

Investigator: Seth Pettie

Some of the most effective algorithms and data structures are based on heuristics that, in the worst case, can lead to suboptimal performance. Perhaps the most well known algorithm in this category is the simplex algorithm for linear programming. In [4] we studied the asymptotic running time of the *pairing heap*, a remarkably simple priority queue supporting all the standard operations. In applications that require the *decreasekey* operation (such as computing shortest paths, minimum weight matchings, and minimum spanning trees) pairing heaps are the priority queue of choice. They are currently incorporated in the GNU C++ library and the LEDA library [citation here]. Despite their empiracle superiority, pairing heaps represent one of the few popular, widely used algorithms whose basic worst case running time remains open.

In earlier analyses [2, 3] it was established that pairing heaps support the deletemin and decreasekey operations in logarithmic time, and insert and meld in constant time, all amortized. Fredman [1] later proved that the running time of decreasekey could, on a particular distribution of operations, take as much as $\Omega(\log \log n)$ time. In [4] we proved that pairing heaps support decreasekey in $2^{O(\sqrt{\log \log n})}$ time without affecting the amortized time of deletemin. (This is the first sub-logarithmic time bound on decreasekey.) One consequence is that in the graph optimization problems cited above, the pairing heap is theoretically no worse than the complicated Fibonacci heap when the input graph is relatively sparse.

References

[1] M. L. Fredman. On the efficiency of pairing heaps and related data structures. *J. ACM*, 46(4):473–501, 1999.

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

[3] J. Iacono. Improved upper bounds for pairing heaps. In *Scandinavian Workshop on Algorithm Theory (SWAT, LNCS 1851*, 2000, pp. 32–43.

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

Dynamic Algorithms for Graph Spanners

Investigator: Surender Baswana

There are a number of applications which require efficient solution of a graph problem in a dynamic environment. In these applications, an initial graph is given, followed by an on-line sequence of updates which can be insertions or deletions of edges. We have to maintain a solution which, depending upon the problem, could either be some graph structure (e.g., minimum spanning tree, matching) or some function defined on the graph (diameter, all-pairs distances) on-line in an efficient manner. The goal of a dynamic graph algorithm is to update the solution efficiently after the dynamic changes, rather than having to re-compute it from scratch each time. There exist efficient dynamic graph algorithms for many important problems – minimum spanning trees, undirected connectivity, Bi-connectivity, transitive closure and all-pairs distances (see [4, 3] and references therein).

We consider the problem of dynamically maintaining graph spanner. A spanner is a sparse subgraph of a given graph that preserves approximate distance between each pair of vertices. In precise words, a t-spanner of a graph $G = (V, E)$, for any $t > 1$ is a subgraph $(V, E_S), E_S \subseteq E$ such that, for any $u, v \in V$, their distance in the subgraph is at most t times their distance in the original graph. The parameter t is called the *stretch* associated with the t-spanner. Spanners (and related structures) are useful in many contexts. They are the basis of space-efficient routing tables that guarantee nearly shortest routes, schemes for simulating synchronized protocols in unsynchronized networks, parallel and distributed algorithms for computing approximate shortest paths. Efficient dynamic algorithms for spanners will find applications in networks and they will also lead to efficient algorithms for maintaining approximate shortest paths in dynamic environments. We present [2] simple and efficient dynamic algorithms for maintaining spanners with essentially optimal (expected) size versus stretch trade-off for any given unweighted graph. The main result is a decremental algorithm that takes expected $O(polylog(n))$ time per edge deletion for maintaining a spanner with arbitrary stretch. This algorithm easily leads to a fully dynamic algorithm with sub-linear (in n) time per edge insertion or deletion. Quite interestingly, we also show that for stretch at most 6, it is possible to maintain a spanner fully dynamically with expected constant time per update. All these algorithms use simple randomization techniques on the top of an existing static algorithm for computing spanners, and achieve drastic improvement over the previous best deterministic dynamic algorithms for spanners [1].

References

[1] G. Ausiello, P. G. Franciosa, and G. F. Italiano. Small stretch spanners on dynamic graphs. In *Proceedings of 13th Annual European Symposium on Algorithms*, 2005, *LNCS 3669*, pp. 532–543. Springer.

[2] S. Baswana. Dynamic algorithms for graph spanners. In *Proceedings of 14th Annual European Symposium on Algorithms (ESA)*, 2006, *LNCS 4168*, pp. 76–87. Springer.

[3] C. Demetrescu and G. F. Italiano. A new approach to dynamic all pairs shortest paths. *Journal of association for computing machinery*, 51:968–992, 2004.

[4] J. Holm, K. de Lichtenberg, and M. Thorup. Poly-logarithmic deterministic fully-dynamic algorithms for connectivity, minimum spanning tree, 2-edge, and biconnectivity. *Journal of association for computing machinery*, 48:723–760, 2001.

Faster Algorithms for Approximate Distance Oracles and All-Pairs Small Stretch Paths

Investigators: Surender Baswana and Kavitha Telikepalli in collaboration with Sandeep Sen

The all-pairs shortest paths (APSP) problem is one of the most fundamental algorithmic graph problems. Efficient algorithms for the APSP problem are very important in several applications. Surprisingly, despite the fundamental importance of the problem, there does not exist any algorithm at present that can solve APSP in truly subcubic time, i.e., $O(n^{3-\epsilon})$ for some $\epsilon > 0$. There exist subcubic algorithms for APSP based on fast matrix multiplication but these algorithms are for small integer edge weights and are notoriously impractical. This has motivated researchers to explore ways to design efficient algorithms for the all-pairs *approximate* shortest paths (APASP) problem. We consider the problem of designing *combinatorial* algorithms (i.e., without using fast matrix multiplication), that are efficient both in terms of space and time required, to compute approximate shortest paths/distances between all pairs of vertices. The final goal of this research is two-fold: to achieve quadratic time bounds for all-pairs approximate distances *and* to obtain space-optimal data structures for this problem in weighted graphs.

An important milestone in the area of algorithms for approximate shortest path is the *approximate distance oracle* by Thorup and Zwick [4]. They show that for any positive integer t, an undirected graph G can be preprocessed to build a data structure that can efficiently report t-approximate distance between any pair of vertices. That is, for any $u, v \in V$, the distance reported is at least $\delta(u, v)$ and at most $t\delta(u, v)$. The remarkable feature of this data structure is that, for $t \geq 3$, it occupies sub-quadratic space, i.e., it does not store all-pairs distances explicitly, and still it can answer any t-approximate distance query in constant time. They named the data structure "approximate distance oracle" because of this feature. Furthermore the trade-off between the stretch t and the size of the data structure is essentially optimal. The construction time of these oracles was sub-cubic (in the number of vertices) and it was posed as an open problem to reduce it to quadratic. As a first step [2], we show that for unweighted graphs, it is possible to construct the approximate distance oracles in worst case quadratic time. However the techniques we used did not seem extendible to weighted graphs. In addition to these approximate distance oracles which are for stretch 3 or more, there also exist algorithms by Cohen and Zwick [3] which work for stretch less than 3. The underlying technique for these algorithms seem fundamentally different from that of approximate distance oracles. As an important contribution [1], we present a single scheme for all-pairs approximate shortest paths for weighted graphs which is arguably better than all the existing ones. This scheme not only can construct approximate distance oracles for weighted graphs in worst case quadratic time, it also improves the running time of the algorithms for APASP with stretch smaller than 3. In particular, our scheme leads to an algorithm that can compute all-pairs *almost* stretch 2 distances in $O(n^2 \log n)$ time.

References

- [1] S. Baswana and T. Kavitha. Faster algorithms for approximate distance oracles and all-pairs small stretch paths. In *In Proceedings of 47th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006, pp. 591–602.
- [2] S. Baswana and S. Sen. Approximate distance oracles for unweighted graphs in expected $o(n^2)$ time. *ACM Transaction on Algorithms*, 2(4):557–577, 2006.

[3] E. Cohen and U. Zwick. All-pairs small stretch paths. *Journal of Algorithms*, 38:335–353, 2001.

[4] M. Thorup and U. Zwick. Approximate distance oracles. *Journal of Association of Computing Machinery*, 52:1–24, 2005.

Sequences Characterizing k-Trees

Investigators: Debapriyo Majumdar and Ingmar Weber in collaboration with Zvi Lotker and N.S. Narayanaswamy

It is easy to prove that a sequence of n integers is the degree sequence of a 1-tree (i.e., an ordinary tree) on n vertices if and only if there are at least two 1's in the sequence, the highest element is at most $n-1$, and the sum of the elements is $2n-2$. In [1] We show the following results towards characterizing the degree sequences of k-trees for $k > 1$.

Firstly, a natural generalization of the statement for 1-trees gives a necessary condition for a sequence to be the degree sequence of a k-tree: the degree sequence of any k-tree must contain at least two k's, the elements of the sequence must be at least k and at most $n-1$, and the sum of the elements must be $k(2n-k-1)$. However, we show that these necessary conditions are not sufficient for any $k > 1$. We construct generic counterexamples of degree sequences which satisfy the necessary conditions and show that such sequences are not degree sequences of any k-tree.

We identify non-trivial sufficient conditions for the degree sequences of 2-trees. For example, we show that in addition to satisfying the necessary conditions, if a sequence contains at least one 3, then it is the degree sequence of a 2-tree.

Using bounds on the partition function $p(n)$ and probabilistic methods we also show that almost all degree sequences satisfying these necessary conditions are degree sequences for some 2-tree.

Finally, we generalize the characterization of degrees of 1-trees in an elegant and counter-intuitive way to yield integer sequences that characterize k-trees, for all k.

References

- [1] Z. Lotker, D. Majumdar, N. Narayanaswamy, and I. Weber. Sequences characterizing k-trees. In D. Z. Chen and D. T. Lee, eds., *Computing and Combinatorics, 12th Annual International Conference, COCOON 2006*, Taipei, Taiwan, 2006, *LNCS 4112*, pp. 216–225. Springer.

Adjacency Queries in Dynamic Sparse Graphs

Investigator: Lukasz Kowalik

We studied fully dynamic graphs, i.e., graphs which change in time by means of inserting and removing edges (it is straightforward to extend our results for the situation when also

vertices may be inserted and removed). Such a setting raises a natural question: how to store the structure of the graph in memory so that some kind of information can be retrieved fast. More specifically, we focus on the most basic sort of information about graph: adjacency. In other words we allow for processing queries of the form *Are vertices u and v adjacent?*. When the graph under consideration is dense, e.g. it has $\Omega(n^2)$ edges (n denotes the number of vertices of the graph) there is a trivial and efficient solution: store an adjacency matrix. Then both the updates and the queries take constant time. However, we consider sparse graphs, i.e., graphs with no dense subgraphs, like for example planar graphs. Then the approach of adjacent matrix is unacceptable because of huge memory requirements compared to the actual size of the graph stored. Hence we focus on data structures of linear space complexity. Then one can use another classic data structure: adjacency lists. Unfortunately, in this case time needed to process a query may be too large, unless there is some bound on the vertices' degrees in the graph.

Brodal and Fagerberg [1] described a very simple linear-size data structure which processes queries in constant worst-case time and performs insertions and deletions in $O(1)$ and $O(\log n)$ amortized time, respectively. We show [2] a complementary result that their data structure can be used to get $O(\log n)$ worst-case time for query, $O(1)$ amortized time for insertions and $O(1)$ worst-case time for deletions. Moreover, our analysis shows that by combining the data structure of Brodal and Fagerberg with efficient deterministic dictionaries one gets $O(\log\log\log n)$ worst-case time bound for queries and deletions and $O(\log\log\log n)$ amortized time for insertions, with size of the data structure still linear. This last result holds even for graphs of arboricity bounded by $O(\log^k n)$, for some constant k.

References

[1] G. S. Brodal and R. Fagerberg. Dynamic representations of sparse graphs. In *Proc. 6th Int. Workshop on Algorithms and Data Structures (WADS'99)*, 1999, *LNCS 1663*, pp. 342–351.

- [2] L. Kowalik. Adjacency queries in dynamic sparse graphs. *Inf. Proc. Lett.*, 2007.

Dualization of Monotone Boolean Functions

Investigators: Khaled Elbassioni in collaboration with Leonid Khachiyan, Endre Boros, Konrad Borys, Vladimir Gurvich, and Kazuhisa Makino

Let $f : \{0,1\}^n \mapsto \{0,1\}$ be a *monotone* Boolean function, defined by its *irredundant disjunctive normal form* (DNF) $f(x) = \bigvee_{F\in\mathcal{F}} \bigwedge_{i\in F} x_i$, where $\mathcal{F}$ is the set of minimal terms of the DNF of f. The dualization problem is to find the corresponding irredundant conjunctive normal form (CNF), i.e. write f in the form $f(x) = \bigwedge_{G\in\mathcal{G}} \bigvee_{i\in G} x_i$, where $\mathcal{G}$ is the set of minimal clauses of the CNF of f. Equivalently, the problem can be stated as of determining if a given pair of a monotone DNF and a monotone CNF represent the same Boolean function (the so-called duality testing), or as of generating all maximal independent sets of a given hypergraph $\mathcal{H} \subseteq 2^V$, i.e. all maximal subsets of V that do not contain any hyperedge of $\mathcal{H}$. Determining the exact complexity of this problem is an outstanding open question, which has already received considerable attention, due to the fact that many other problems, of the enumeration type, can be shown to be polynomially equivalent with it. The best algorithm for solving this problem, known since 1996, runs in time quasi-polynomial

in $(m+k)^{o(\log(m+k))}$, where $m = |\mathcal{F}|$ and $k = |\mathcal{G}|$. In [2], we improve this best bound in two ways: First we show that the exponent in the bound can be made to *depend only* on either m or k, which is of particular interest if we consider the generation of maximal independent sets of a given hypergraph $\mathcal{H}$, since typically the size of the output hypergraph is much larger than the size of the input $|\mathcal{H}|$. Second, we show that the computation of each new maximal independent set can be achieved with a recursion tree of *polylogarithmic* depth, implying that the the duality testing problem can be solved in polylogarithmic-time on a quasi-polynomial number of processors. When all the terms of the input DNF are of bounded size, we show in [4] that the corresponding CNF can be found in $polylog(n, m, k)$ parallel time using $poly(n, m, k)$ number of processors. Some generalizations of this problem to products of lattices and to matroids, with applications to enumeration problems in data mining and reliability theory, are considered in [1] and [3].

References

- [1] K. Elbassioni. Finding all minimal infrequent multi-dimensional intervals. In J. R. Correa, A. Hevia, and M. A. Kiwi, eds., *LATIN 2006: Theoretical Informatics, 7th Latin American Symposium*, Valdivia, Chile, 2006, vol. 3887, pp. 423–434. Springer.
- [2] K. Elbassioni. On the complexity of the multiplication method for monotone cnf/dnf dualization. In Y. Azar and T. Erlebach, eds., *Algorithms - ESA 2006, 14th Annual European Symposium*, Zürich, Switzerland, 2006, *LNCS 4168*, pp. 340–351. Springer.
- [3] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, V. Gurvich, and K. Makino. Enumerating spanning and connected subsets in graphs and matroids. In Y. Azar and T. Erlebach, eds., *Algorithms - ESA 2006, 14th Annual European Symposium*, Zürich, Switzerland, 2006, *LNCS 4168*, pp. 444–455. Springer.
- [4] L. Khachiyan, E. Boros, K. Elbassioni, and V. Gurvich. A new algorithm for the hypergraph transversal problem. In L. Wang, ed., *Computing and combinatorics, 11th Annual International Conference, COCOON 2005*, Kunming, China, August 2005, *LNCS 3595*, pp. 767–776. Springer.

12.4 Combinatorial Optimization

Coordinators: Ernst Althaus and Kurt Mehlhorn

Many real world applications are naturally formulated as combinatorial optimization problems, i.e. problems of finding the best solution 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, e.g. generic methods for integer programming problems, scheduling problems, shortest path, and problems from industrial cooperations and from bioinformatics.

12.4.1 Integer Programming

Integer programming is a very important method to tackle combinatorial optimization problems as many problems can be formalized within this framework. We work on methods to

improve the efficiency of algorithms for integer programming problems and on the theoretical analysis of them.

0/1 Vertex and Facet Enumeration with BDDs

Investigators: Markus Behle and Friedrich Eisenbrand

This research was partially developed in NWG2 and can be found in Section 19.2.1.

Smoothed Analysis of Multicriteria Integer Optimization Problems

Investigators: René Beier in collaboration with Heiko Röglin and Berthold Vöcking

A well established heuristic approach for solving various bicriteria optimization problems is to enumerate the set of Pareto optimal solutions. These solutions are optimal compromises of the criteria in the sense that any improvement of one criterion implies an impairment to the other. The heuristics following this principle are often successful in practice. Their running time, however, depends on the number of enumerated solutions, which can be exponential in the worst case.

In [1] we prove an almost tight bound on the expected number of Pareto optimal solutions for general bicriteria integer optimization problems in the framework of smoothed analysis. We requires that one objective function is linear of the form $p^{\mathrm{T}}x$. In our semi-random input model we assume that an adversary first specifies an input. Subsequently, only the coefficients of the linear objective function are perturbed, leaving the underlying combinatorial structure of the problem untouched. We perturb by adding an independent Gaussian random variable with mean 0 and standard deviation σ to each coefficient p_i. The standard deviation σ can be seen as a parameter measuring how close the analysis is to a worst-case analysis: The smaller σ is chosen, the smaller is the influence of the perturbation and, hence, the closer is the analysis to a worst-case analysis. Our probabilistic analysis is not restricted to Gaussian perturbations but covers arbitrary probability distributions with a bounded density function and a finite absolute mean value.

Applying our findings to the 0/1 knapsack problem, we can improve on previous results and obtain tight bounds on the expected number of Pareto optimal knapsack fillings confirming a conjecture from an experimental study [2]. Consequently we also obtain tight polynomial bounds on the expected running time of the Nemhauser/Ullmann heuristic for the 0/1 knapsack problem. Furthermore, we can significantly improve the known results on the running time of heuristics for the bounded knapsack problem and for the bicriteria shortest path problem.

Another common approach to tackle multi-criteria optimization problems chooses one criterion to be the only objective function and bounds the other criteria by appropriate constants. Using a similar semi-random input model, we could prove in [3] that for binary programs the slack of the optimal solution with respect to the additional constraints is usually large, i.e., polynomially bounded in σ/n, where n denotes the number of variables. A similar bound holds for the smallest negative slack over all solutions with better objective function value than the optimum. Hence, one can round the coefficients of the linear constrains up to a certain degree without changing the optimal solution. Using an adaptive

rounding scheme, we can show that a binary optimization problem in this form has polynomial smoothed complexity if and only if there exists a pseudo-polynomial (w. r. t. the coefficients of the additional constraints) time algorithm for solving the problem. The term *polynomial smoothed complexity* is defined analogously to the way polynomial complexity is defined in average-case complexity theory, adding the requirement that the running time should be polynomially bounded not only in the input size but also in $1/\sigma$. Using results from [1] we can generalize this result from binary variables to integer variables that take values from a bounded domain.

References

- [1] R. Beier, H. Röglin, and B. Vöcking. The smoothed number of pareto optimal solutions in bicriteria integer optimization. In *12th Conference on Integer Programming and Combinatorial Optimization*, Ithaca, USA, 2007, pp. 000–999. Springer. To Appear.
- [2] R. Beier and B. Vöcking. An experimental study of random knapsack problems. *Algorithmica*, 45(1):121–136, February 2006.
- [3] R. Beier and B. Vöcking. Typical properties of winners and losers in discrete optimization. *SIAM Journal on Computing*, 35(4):855–881, February 2006.

12.4.2 Approximation Algorithms

As many combinatorial optimization problems are NP-hard, there is no hope for polynomial algorithms that compute an optimal solution. Hence we are interested in approximation algorithms, i.e. algorithms that run in polynomial time and find a solution that is provably not far from the optimal solution.

Simultaneous Matchings: Hardness and Approximation

Investigators: Khaled Elbassioni, Irit Katriel, Martin Kutz, and Meena Mahajan

Given a bipartite graph $G = (X \cup D, E \subseteq X \times D)$, an *$X$-perfect matching* is a matching in G that covers every node in X. In this work [1], we study the following generalization of the X-perfect matching problem, which has applications in constraint programming: Given a bipartite graph as above and a collection $\mathcal{F} \subseteq 2^X$ of k subsets of X, find a subset $M \subseteq E$ of the edges such that for each $C \in \mathcal{F}$, the edge set $M \cap (C \times D)$ is a C-perfect matching in G (or report that no such set exists). We show that the decision problem is NP-complete and that the corresponding optimization problem is in APX when $k = O(1)$ and even APX-complete already for $k = 2$. On the positive side, we show that a $2/(k+1)$-approximation can be found in $poly(k, |X \cup D|)$ time. We show also that such an approximation M can be found in time that each vertex in D has degree at most 2 in M.

References

- [1] K. Elbassioni, I. Katriel, M. Kutz, and M. Mahajan. Simultaneous matchings. In X. Deng and D. Du, eds., *Algorithms and computation, 16th International Symposium, ISAAC 2005*, Sanya, Hainan, China, 2005, *LNCS 3827*, pp. 106–115. Springer.

Rounding SDP Relaxations for Vertex Ordering Problems

Investigator: Alantha Newman

An important class of optimization problems are *vertex ordering* problems. In such problems, the objective is to arrange the vertices of a given graph on a line subject to specified constraints so as to maximize a given objective function. Many simple-to-state NP-hard optimization problems such as the *linear ordering problem*, the *traveling salesman problem*, and the *minimum bandwidth problem* fall into the class of vertex ordering problems. These optimization problems can often be viewed as assignment problems, which can be formulated as quadratic integer programs based on assignment constraints. Thus, a possible general approach to approximating such problems is to round the corresponding semidefinite programming (SDP) relaxations. However, an obstacle to this approach is our relatively sparse knowledge of techniques for rounding SDPs that model assignment constraints over large domains. In fact, there has been much recent interest in optimization problems that can be classified as assignment problems over large domains, such as Unique Games, upon which the Unique Games Conjecture is based. Several algorithms for this problem are based on rounding semidefinite programs that are modeled on assignment constraints, e.g. [2]. In particular, SDP-based techniques have been successfully applied to a range of assignment problems over small domains (e.g. of size two or three [3, 4]), using essentially the same strategy, which has not been rigorously analyzed for smaller domains of size more than three and which does not seem to be a promising approach in general to problems on large domains (e.g. of size $\Omega(n)$).

We study new rounding techniques for semidefinite programs over large domains. Specifically, we have conducted computational experiments in which we apply these techniques to the linear ordering problem, in an attempt to find good heuristics for approximating this problem [5]. (Given a directed graph, the goal of the linear ordering problem is to find an ordering of the vertices that maximizes the weight of the forward edges.) Additionally, we have conducted computational experiments in which we apply these techniques to variants of the classical constraint satisfaction problem known as *linear equations* mod p, in which we are given equations of the form $x_i - x_j \equiv c_k \bmod p$ and the goal is to assign the variables $\{x_i\}$ values from the range 0 through $p-1$ so as to satisfy as many of the given equations as possible. In the latter case, our experimental results show that we can obtain non-trivial approximations for the variants that we consider. These variants generalize the *maximum cut problem*, suggesting that it is necessary to use a semidefinite program to obtain a non-trivial approximation. Furthermore, previously used techniques [1] can easily be shown to perform poorly when applied to these particular variants. We are currently working on proving some basic geometric conjectures that would explain the positive behavior of our algorithms demonstrated by our experiments.

References

[1] G. Andersson, L. Engebretsen, and J. Håstad. A new way to use semidefinite programming with applications to linear equations mod p. *Journal of Algorithms*, 39:162–204, 2001.

[2] M. Charikar, K. Makarychev, and Y. Makarychev. Near-optimal algorithms for unique games. In *Proceedings of the 38th Annual Symposium on the Theory of Computing (STOC)*, Seattle, 2006.

[3] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995.

[4] M. X. Goemans and D. P. Williamson. Approximation algorithms for MAX-3-CUT and other problems via complex semidefinite programming. *STOC 2001 Special Issue of Journal of Computer and System Sciences*, 68:442–470, 2004.

- [5] A. Newman. An experimental study of SDP-based heuristics for approximating the maximum acyclic subgraph. 2006.

Maximum Asymmetric TSP with Triangle Inequality

Investigators: Lukasz Kowalik and Marcin Mucha

The Traveling Salesman Problem and its variants are among the most intensively researched problems in computer science and arise in a variety of applications. In its classical version, given a set of vertices V and a symmetric weight function $w : V^2 \to \mathbb{R}$ one has to find a Hamiltonian cycle of minimum weight. This problem is probably the most widely known example of an inapproximable NP-hard problem. However, there is a lot of research on approximation of several natural variants of TSP. These variants are still NP-hard, but allow approximation. One of the most important problems in this category is the maximization version (maxTSP for short), where w is assumed to have only nonnegative values (otherwise minTSP would reduce to it). There are several variants of maxTSP, for example the weight function can be symmetric or asymmetric, it can satisfy the triangle inequality or not, etc. (For some results on maxTSP variants see e.g. [2, 3, 5, 7]).

Here, we are concerned with the variant, where the weight function is asymmetric (in other words, the graph is directed) and satisfies the triangle inequality. This variant is often called *the semimetric maxTSP*.

The first approximation algorithm for this problem was proposed by Kostochka and Serdyukov [8] in 1985 and had approximation ratio of 3/4. Quite recently, Kaplan, Lewenstein, Shafrir and Sviridenko [4] provided a very general and powerful framework for approximating asymmetric TSP variants and gave improved approximation ratios for 3 different problems: $\frac{4}{3}\log_3 n$ for semimetric minTSP, $\frac{10}{13}$ for semimetric maxTSP and $\frac{2}{3}$ for asymmetric maxTSP. Using a more elaborate analysis and slightly different algorithm, Feige and Singh [1] were later able to achieve an approximation factor of $\frac{2}{3}\log_2 n$ for semimetric minTSP.

We show in [6], that in the case of semimetric maxTSP the ideas of Kaplan et al. can be combined with a new patching procedure yielding an approximation factor of 35/44.

References

[1] U. Feige and M. Singh. Improved approximation ratios for traveling salesperson tours and paths in directed graphs, 2006. Manuscript.

[2] R. Hassin and S. Rubinstein. Better approximations for max tsp. *Inf. Process. Lett.*, 75(4):181–186, 2000.

[3] R. Hassin and S. Rubinstein. A 7/8-approximation algorithm for metric max tsp. *Inf. Process. Lett.*, 81(5):247–251, 2002.

[4] H. Kaplan, M. Lewenstein, N. Shafrir, and M. Sviridenko. Approximation algorithms for asymmetric tsp by decomposing directed regular multigraphs. *J. ACM*, 52(4):602–626, 2005.

[5] S. R. Kosaraju, J. K. Park, and C. Stein. Long tours and short superstrings (preliminary version). In *FOCS*, 1994, pp. 166–177.

- [6] L. Kowalik and M. Mucha. 35/44-approximation for asymmetric maxTSP with triangle inequality. *submitted to WADS07*, 2007.

[7] M. Lewenstein and M. Sviridenko. A 5/8 approximation algorithm for the maximum asymmetric tsp. *SIAM J. Discrete Math.*, 17(2):237–248, 2003.

[8] A. I. Serdyukov. The traveling salesman problem of the maximum (in russian). *Upravlyaemye Sistemy*, 25:80–86, 1984.

Approximation Algorithms for Maximum Symmetric TSP

Investigator: Katarzyna Paluch

We consider the symmetric Maximum Traveling Salesman Problem, in which we are given a complete undirected graph $G = (V, E)$ with nonnegative weights on the edges and we wish to find a traveling salesman tour of maximal weight. A tour is a simple cycle that contains each vertex of G. We give a 7/9 approximation algorithm. The algorithm is fast, deterministic and combinatorial. The best previous bounds on the approximation were 3/4 by Serdyukov (1984), recently improved to 61/81 by Chen et.al. in 2005. We view a traveling salesman tour as a cycle cover and use the fact that the cycle cover of maximum weight is the upper bound on the weight of the optimal tour. Cycle covers, in turn, are 2-matchings and using the theory about matchings we construct a better upper bound for Max TSP [1].

References

- [1] K. Paluch. A new approximation algorithm for the maximum traveling salesman problem. 2007.

Approximation Algorithms for Multidimensional Rectangle Tiling

Investigator: Katarzyna Paluch

We consider the following tiling problem: Given a d-dimensional array A of size n in each dimension, containing non-negative numbers and a positive integer p, partition the array A into at most p disjoint rectangular subarrays called *rectangles* so as to minimize the maximum weight of any rectangle. The weight of a subarray is the sum of its elements. The problem has applications in databases, load balancing and video compression.

The multidimensional version was first considered by Smith and Suri in [4], where they give an algorithm with approximation ratio $\frac{d+3}{2}$, that runs in time $O(n^d + p \log n^d)$ and the constant is of the order of $d!$. Next, Sharp in [3] gave a $(d^2 + 2d - 1)/(2d - 1)$-approximation algorithm that runs in time $O(n^d + 2^d p \log n^d)$.

We give a $\frac{d+2}{2}$-approximation algorithm that runs in time $O(n^d + 2^d p \log n^d)$ [2]. Additionally, this algorithm is tight with regard to the only known and used lower bound so far.

The general approach has a similar spirit as that in [1]. We also classify the arrays and subarrays into types. In the multidimensional case, however, there are many kinds of subarrays with a short type (having length 2) that are difficult to partition (whereas in a two dimensional case there is only one kind of such subarrays). As previously, we also have to consider arbitrarily large subarrays i.e. having arbitrarily long type. Fortunately subarrays that are difficult to partition display a regular structure that can be handled by appropriate linear programs. Curiously, linear programs describing large difficult subarrays disintegrate into small linear programs that can be treated independently and in this respect they are much easier to analyze than linear programs describing large difficult subarrays in a two dimensional version, where they cannot be decomposed into small linear programs.

References

[1] K. Paluch. A 2(1/8)-approximation algorithm for rectangle tiling. In *ICALP*, 2004, pp. 1054–1065.

- [2] K. Paluch. A new approximation algorithm for multidimensional rectangle tiling. In T. Asano, ed., *Algorithms and Computation, 17th International Symposium, ISAAC 2006*, Kolkata, India, 2006, pp. 712–721. Springer.

[3] J. Sharp. Tiling multi-dimensional arrays. In *Fundamentals of Computation Theory*, 1999, pp. 500–511.

[4] A. Smith and S. Suri. Rectangular tiling in multi-dimensional arrays. *Journal of Algorithms*, 37(2):451–467, 1999.

Scheduling on Related Machines

Investigator: Annamária Kovács

We apply the standard scheduling notation $Q||C_{\max}$ for the classic problem of allocating n jobs to m machines of different speeds so that the maximum finish time over all machines (makespan) is minimized.

In the course of obtaining the 3-approximation monotone algorithm for $Q||C_{\max}$ [4] (see also Section 12.3.4), as a core result we proved that the greedy list scheduling heuristic *Largest Processing Time first* (LPT) is monotone, and yields 3/2-approximation if machine speeds are all integer powers of two (*2-divisible* machines). Motivated by the wish to reduce the above bound of 3/2 (and consequently the bound 3), in the paper [5] we further considered $Q||C_{\max}$ in the special case of 2-divisible machines. We provided an analysis of the ratio Lpt/Opt. We showed that in the worst case $1.3673 < Lpt/Opt < 1.4$. Additionally, we proved another lower bound of $955/699 > (\sqrt{3}+1)/2$ for arbitrary tie breaking of LPT.

Moreover, as a side result, we obtained a tight bound of $(\sqrt{3}+1)/2 \approx 1.3660$, for another special speed vector: in the case of *'one fast machine'*, i.e., when $m-1$ machine speeds are equal, and there is only one faster machine. This special case has been widely studied in the literature. The best previous lower and upper bounds of 4/3, and $3/2 - 1/2m$ were due to Gonzalez, Ibarra and Sahni [3], who conjectured the lower bound 4/3 to be tight.

Recently we have investigated the approximation bound of LPT for *arbitrary* machine speed vectors. The best previously known bounds originate from more than 20 years back: Dobson [1], and independently Friesen [2] showed that the worst case ratio of LPT is in the interval $(1.512, 19/12)$, and in $(1.52, 1.67)$, respectively. We could tighten the upper bound

of $19/12 \approx 1.5833$ to $1 + \sqrt{3}/3 \approx 1.5773$, and provide a scheme for a job-exchanging process, which, repeated any number of times, gradually increases the lower bound. However, the limit of these increased lower bounds is conjectured to be below 1.53. Although the improvement might seem minor, we consider the structure of (potential) instances providing higher bounds more systematically than former works. For the new upper bound, this systematic method together with a new idea, facilitated a proof that is surprisingly simple, and more consistent than the old proof for 19/12.

References

[1] G. Dobson. Scheduling independent tasks on uniform processors. *SIAM Journal on Computing*, 13(4):705–716, 1984.

[2] D. Friesen. Tighter bounds for LPT scheduling on uniform processors. *SIAM Journal on Computing*, 16(3):554–560, 1987.

[3] T. Gonzalez, O. Ibarra, and S. Sahni. Bounds for LPT schedules on uniform processors. *SIAM Journal on Computing*, 6(1):155–166, 1977.

- [4] A. Kovács. Polynomial time preemptive sum-multicoloring on paths. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, eds., *Automata, languages and programming, 32nd International Colloquim, ICALP 2005*, Lisbon, Portugal, July 2005, *LNCS 3580*, pp. 840–852. Springer.

- [5] A. Kovács. Tighter approximation bounds for LPT scheduling in two special cases. In T. Calamoneri, I. Finocchi, and G. F. Italiano, eds., *Algorithms and Complexity, 6th Italian Conference, CIAC 2006*, Rome, Italy, May 2006, *LNCS 3998*, pp. 187–198. Springer.

On the Value of Preemption in Scheduling

Investigators: René Sitters in collaboration with Gil Shallom, Yair Bartal, and Stefano Leonardi

Online scheduling algorithms can achieve optimal performance if preemption of tasks is allowed. A classic example is the shortest remaining processing time (SRPT) algorithm which is optimal for minimum flow time scheduling. Optimality holds under the assumption that preemption is costless. But in real systems, preemption usually has significant overhead. We develop online scheduling algorithms for a costly preemption model and show that, within this model, our algorithms perform best possible with respect to worst-case analysis [1].

References

- [1] Y. Bartal, S. Leonardi, G. Shallom, and R. Sitters. On the value of preemption in scheduling. In J. Diaz, K. Jansen, J. D. P. Rolim, and U. Zwick, eds., *9th Workshop on Approximation Algorithms for Combinatorial Optimization Problems, APPROX*, Barcelona, Spain, August 2006, *LNCS 4110*, pp. 39–48. Springer.

12.4.3 Partner Group on Approximation Algorithms

Coordinator: Naveen Garg

The Algorithms Group at IIT Delhi, India was designated a Partner-Group of the Max Planck Institute for Informatics for research in the area of "Approximation Algorithms". The group started its operations in July 2005.

Minimizing Flow Time on Related Machines

Investigators: Naveen Garg and Amit Kumar

A well-studied setting in the scheduling literature is one where we have multiple machines of differing speeds. This is commonly referred to as the related machine scenario. All machines are equally capable so that a job can be scheduled on any machine. The only distinction is that a machine of speed twice that of another machine would take half the time to finish the same job.

The jobs arrive over time and have to be scheduled so that the total flow time is minimized. The flow time of a job is the difference between its completion and release times and is equal to the total time that the job is waiting or being processed. We permit preemption, so that a job can be stopped even before it is completed and resumed later. However, we do not permit migration, i.e. the job cannot be resumed, after preemption, on another machine. Thus we are looking for a preemptive, non-migratory schedule on m related machines which minimizes the total flow time. In the three field notation of scheduling problems this is denoted by $Q|r_j, pmtn|\sum_j (C_j - r_j)$.

When all machines have the same speed – the setting of parallel machines – the problem is well-studied. Leonardi and Raz [3] showed that the Shortest-Remaining-Processing-Time (SRPT) rule gives a schedule which is $O(\min(\log P, \log(n/m)))$-competitive, where P is the ratio of the maximum processing time to the minimum processing time and n is the number of jobs. They also established an $\Omega(\log P)$ lower bound on the competitiveness of any randomized online algorithm.

In [2] we give the first on-line poly-logarithmic competitive algorithm for this problem. More specifically, we give an $O(\log^2 P \log S)$-competitive algorithm, where P is the ratio of the biggest and the smallest processing time of a job, and S is the ratio of the highest and the smallest speed of a machine. Our algorithm also has the nice property that it is non-migratory and is based on the concept of making jobs wait for a long enough time before scheduling them on slow machines.

In [1] we develop a new linear programming approach to minimizing flow time on related machines. Our LP is a natural extension of the preemptive time-indexed formulation introduced in the context of single machine scheduling approximation algorithms. We show a way of rounding the LP solution to obtain a non-migratory schedule whose flow time is at most $O(\log P)$ times the objective value of the LP solution. This gives an offline, $O(\log P)$-approximation algorithm for this problem.

We do not know how to solve this LP in an online manner. However, we can modify the linear program so that an optimum solution to the modified LP can be found in an online manner. Our procedure for rounding the LP solution into a non-migratory schedule continues to apply even in this online setting. Finally we show that the optimum solution to our modified LP is at most $O(\log P)$ times the flow time of the best schedule. This implies an $O(\log^2 P)$-competitive online algorithm for minimizing flow time on related machines.

References

- [1] N. Garg and A. Kumar. Better algorithms for minimizing average flow-time on related machines. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part I*, 2006, *LNCS 4051*, pp. 181–190. Springer.
- [2] N. Garg and A. Kumar. Minimizing average flow time on related machines. In *STOC '06: Proceedings of the thirty-eighth Annual ACM Symposium on Theory of Computing*, New York, NY, USA, 2006, pp. 730–738. ACM Press.

[3] S. Leonardi and D. Raz. Approximating total flow time on parallel machines. *Proceedings of the twenty-ninth Annual ACM Symposium on Theory of Computing*, pp. 110–119, 1997.

Lagrangian Relaxation in Algorithm Design

Investigators: Garima Batra, Naveen Garg, Garima Gupta, and Parul Jain

Many algorithms in Combinatorial Optimization and Computer Science require the solution of a linear program. These Linear Programs (LP) capture the combinatorial structure of the problem as a set of variables and constraints. While one can always solve these LPs by using one of the generic algorithms for Linear Programming these techniques do not utilize the special structure of such linear programs. With this in mind researchers have developed Lagrangian relaxation algorithms in which the combinatorial structure of the linear program (the set of "easy constraints") are used to define a polytope such that efficient algorithms to optimize linear functions of the polytope are available. The "hard constraints" are relaxed but a penalty is imposed on their violation (Lagrangian relaxation).

Our group has been working on developing new algorithms using such an approach and also on obtaining fast implementations of such algorithms.

Online Packing

An integer packing problem is a problem of the form

$$\left\{ \max \sum_{j=1}^{n} c_j x_j | \sum_{j=1}^{n} a_{ij} x_j \leq b_i, 1 \leq i \leq m \right\}$$

where x_j is a non-negative integer and all entries a_{ij}, b_i, c_j are non-negative. In the Online version of this problem, we get the columns of A and the corresponding coefficient in the objective, online, and have to choose an appropriate value for x_j which cannot be modified subsequently.

A special case of this problem, that of routing virtual circuits, was first considered by Awerbuch, Azar and Plotkin [1]. They showed a $O(\log mF)$ competitive algorithm under the assumption that the minimum edge capacity was at least $\Omega(\log mF)$ times the maximum demand requested, where F is an upper bound on the ratio of the demand to the profit of a connection.

Let $B = \min_{i,j} b_i/a_{ij}$. We show a very simple online algorithm which gives a solution x, which for all i, satisfies $\sum_j a_{ij}x_j \leq b_i(\log_L(4m\text{opt}/\alpha) + 1/B)$ and has value at least $\text{opt}/4B(L^{1/B}-1)$. Here α is a lower bound on the optimum value, opt, and L is any constant larger than 1. Our algorithm can be modified to yield a fractional solution, x, which for all i, satisfies $\sum_j a_{ij}x_j \leq b_i \log_L(4m\text{opt}/\alpha)$ and has value at least $\text{opt}/4\ln L$. This matches the bound obtained by Buchbinder and Naor [3] for this problem.

Heuristic Improvements for Maximum Multicommodity Flow

We consider the problem of computing the maximum multicommodity flow in a given undirected capacitated network with k source-sink pairs. The algorithm with the best asymptotic running time for this problem is due to Fleischer [4] and computes a $1+\omega$ approximation to the maximum multicommodity flow in time $\mathrm{O}((m/\omega)^2 \log m)$. A closely related quantity is the minimum multicut in the graph which is NP-hard to compute and for which there is an $\mathrm{O}(\log k)$-approximation algorithm.

In [2] we show how to implement these algorithms so as to reduce the running time in practice. One novelty of our approach lies in our use of the minimum multicut to speed up the computation. This has the added advantage that we get a multicut which is much better than the one which would have been obtained by following the algorithm of Garg-Vazirani-Yannakakis.

References

[1] B. Awerbuch, Y. Azar, and S. Plotkin. Throughput-competitive on-line routing. *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pp. 32–40, 1993.

- [2] G. Batra, N. Garg, and G. Gupta. Heuristic improvements for computing maximum multicommodity flow and minimum multicut. In *Algorithms - ESA 2005, 13th Annual European Symposium, Palma de Mallorca, Spain, October 3-6, 2005, Proceedings*, 2005, *LNCS 3669*, pp. 35–46. Springer.

[3] N. Buchbinder and J. Naor. Online primal-dual algorithms for covering and packing problems. *13th Annual European Symposium on Algorithms-ESA 2005*, 2005.

[4] L. Fleischer. Approximating fractional multicommodity flow independent of the number of commodities. In *Symposium on Foundations of Computer Science (FOCS)*, 1999, pp. 24–31. IEEE Computer Society.

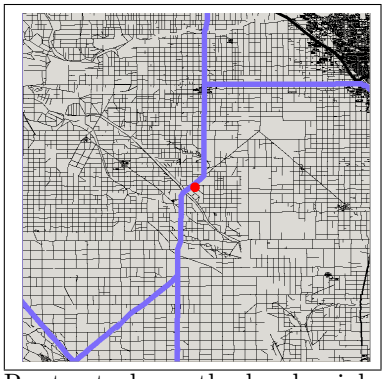
12.4.4 Applied Optimization

D1 developed and implemented efficient algorithms for several real world problems. We give new methods for the well known shortest path problem and for two applications from industrial cooperations.

Ultrafast Shortest-Path Queries via Transit Nodes

Investigators: Holger Bast, Stefan Funke and Domagoj Matijevic in collaboration with Peter Sanders and Dominik Schultes

Computing an optimal route in a road network between specified source and target nodes is one of the showpieces of real-world applications of algorithmics. Besides the omnipresent application of car navigation systems and Internet route planners, even faster route planning is needed for massive traffic simulation and optimization in logistics systems.



Routes to leave the local neighborhood of a small village south of Fresno (USA).

The classical algorithm for route planning – Dijkstra's algorithm [3] – iteratively visits all nodes that are closer to the source node than the target node before reaching the target. On road networks for a subcontinent like Western Europe or the USA, this takes about ten seconds on a state-of-the-art workstation. Since this is too slow for many applications, commercial systems use heuristics that do not guarantee optimal routes. Therefore, there has been considerable interest in speedup techniques for computing *optimal* routes. The most promising approach is to preprocess information exploiting special properties of road networks. Previous approaches include goal directed search (e.g. [8, 5, 4]), and exploiting that road networks can be partitioned into small pieces by removing a small number of nodes (e.g. [6, 7]).

Our approach, coined *transit node routing*, is based on on a very simple, intuitive insight into the structure of road networks: When you drive to somewhere 'far away', you will leave your local neighborhood via one of only very few routes. Based on this observation we can identify a relatively small set of nodes what we call *transit nodes*, about 10 000 for the European road network, with the property that for every pair of nodes that are 'not too close' to each other, the quickest path between them passes through *at least one* of these transit nodes. Second, for every node, the set of transit nodes encountered first when going far – we call these *access nodes* – is small (about ten). The idea behind transit node routing is to precompute travel times between transit nodes and travel times from all nodes to their access nodes. Together with an effective notion of 'not too close', this allows most queries to be answered using a few table lookups. Our fastest query times are about 6 μs, which is by orders of magnitudes faster than all previously known results. See [1] and [2].

References

- [1] H. Bast, S. Funke, and D. Matijevic. Transit: Ultrafast shortest-path queries with linear-time preprocessing. In C. Demetrescu, A. Goldberg, and D. Johnson, eds., *9th DIMACS Implementation Challenge — Shortest Path*, Piscataway, New Jersey, 2006. DIMACS.

- [2] H. Bast, S. Funke, D. Matijevic, P. Sanders, and D. Schultes. In transit to constant time shortest-path queries in road networks. In D. Applegate and G. Brodal, eds., *9th Workshop on Algorithm Enginneering and Experiments (ALENEX'07)*, New Orleans, USA, 2007. SIAM.

[3] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[4] A. Goldberg, H. Kaplan, and R. Werneck. Reach for a*: Efficient point-to-point shortest path algorithms. In *Workshop on Algorithm Engineering & Experiments*, Miami, 2006, pp. 129–143.

[5] U. Lauther. An extremely fast, exact algorithm for finding shortest paths in static networks with geographical background. In *Proceedings Münster GI-Days*, 2004.

[6] F. Schulz, D. Wagner, and C. D. Zaroliagis. Using multi-level graphs for timetable information. In *4th Workshop on Algorithm Engineering and Experiments*, 2002, *LNCS 2409*, pp. 43–59. Springer.

[7] M. Thorup. Compact oracles for reachability and approximate distances in planar digraphs. *J. ACM*, 51(6):993–1024, 2004.

[8] D. Wagner and T. Willhalm. Geometric speed-up techniques for finding shortest paths in large sparse graphs. In *11th European Symposium on Algorithms*, 2003, *LNCS 2832*, pp. 776–787. Springer.

Trunk Packing

Investigators: Ernst Althaus, Friedrich Eisenbrand, Stefan Funke, Andreas Karrenbauer, Joachim Reichel, Jens Rieskamp and Kai Werth in collaboration with Tobias Baumann and Elmar Schömer

In an industry project with a German car manufacturer we are faced with the challenge of placing a maximum number of uniform rigid rectangular boxes in the interior of a car trunk. The problem is of practical importance due to a European industry norm [3] which requires car manufacturers to state the trunk volume according to this measure.

Based on a result by Fowler, Paterson and Tanimoto [6], we show the NP-completeness of our problem. Other results in the area of industrial packing problems suggest that exact solutions can be computed only for very small problem instances. For example, Daniels and Milenkovic [7, 2] consider the problem of minimizing cloth utilization when cutting out a small number of pieces from a roll of stock material. Aardal and Verweij [11] consider the problem of labelling points on a map with pairwise disjoint rectangles.

We pursue two different approaches, namely the discrete (combinatoric) [5, 8] and the continuous [4, 9] approach.

To reduce the complexity of the problem, we perform two discretization steps that are suggested by the manual packings used so far. First, we approximate the interior of the trunk by a three-dimensional cubic grid. Second, we discretize the box placements by demanding that all boxes have to be aligned with this grid. Now our problem can be restated as a maximum stable set for the so-called *conflict graph*.

We use integer linear programming techniques to solve the stable set problem. Although we are using tight formulations (e.g. clique inequalities, violated lifted odd cycles), our problem instances are too big to obtain a good solution in reasonable time. Still this exact algorithm is useful for solving smaller subproblems. Therefore we investigate heuristics based on LP rounding and on the Reactive Local Search algorithm developed by Battiti and Protasi [1]. We further reduce the complexity of the problem instances by using simple geometry guided heuristics, which provide a dense packing of the center region. The algorithms above are then applied to the remaining space.

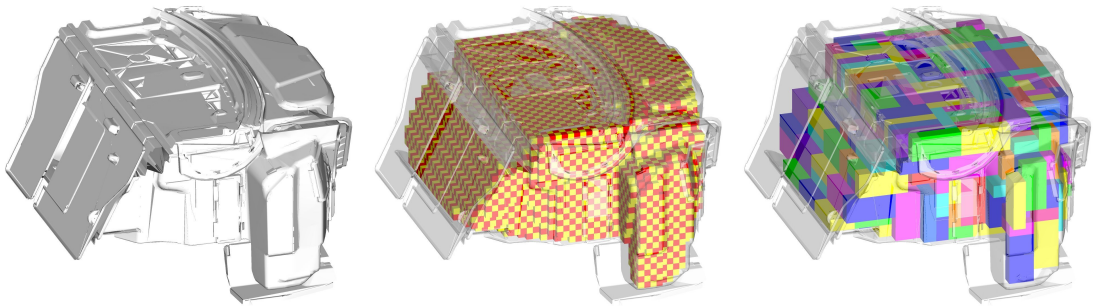


Figure 12.6: CAD model of the trunk (left) and discretization of its interior by a cubic grid (middle); computed packing of boxes (right)

After 24 hours running time, for most instances, we obtain solutions that reach 99% of the volume of a manual packing constructed by an experienced engineer. Though, when evaluating this system in the industrial production environment, some instances with unsatisfying results showed up. The remaining gap is mostly due to the constraints imposed by the chosen grid. While the orientation of the grid is suitable for most regions of the trunk, there are areas where a different orientation is necessary to avoid wasting space. This happens in particular along curved parts of a trunk, see for example in Fig. 12.7, where the restriction to one cubical grid system wastes a lot of volume along the curved lid (left picture) compared to a solution with arbitrary rotations (right picture).

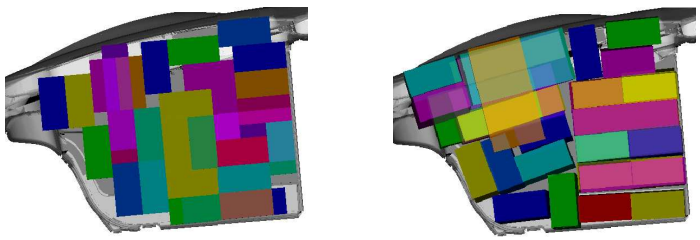


Figure 12.7: Curved lid

Therefore we pursue a continuous approach allowing arbitrary orientations and placements of boxes. Such a continuous model leads to a very high-dimensional global optimization problem for which standard methods like *Simulated Annealing* are typically used. Unfortunately, applying these techniques in a straightforward manner yields solutions far worse than the discretization algorithm. Only by combining both techniques we were able to obtain the industry-strength system.

Roughly speaking, we eliminate the heating process of standard Simulated Annealing by using a solution from the discrete approach as starting configuration. In our algorithm, we iteratively apply a sequence consisting of (a) a special creation procedure for new boxes, (b) a relaxation period by a Monte Carlo simulation, and (c) a randomly triggered destruction of the "worst" box. Due to physical analogies, we call our method *Specialized Grand Canonical*

Simulated Annealing.

We use a self-adaptive scheme to control the numerous parameters of the SGCSA approach. Its performance can be further enhanced by considering specialized moves during the relaxation phase: For example, instead of randomly moving a single box per iteration, a set of boxes, e.g., several boxes in a row, are considered.

The combination of discrete and continuous approach allows to non-interactively compute solutions that meet the strict quality requirements of our industrial partner. Our software package enables car manufacturers to reliably estimate the volume of car trunks at an early stage of the design process.

In a follow-up project we are concerned with a different standard: In the United States, the volume of a car trunk has to measured according to SAE J1100 [10], which defines a *standard luggage set* (modeled as boxes of different sizes) that has to be packed into the trunk. The main difference that can be used algorithmically is that the boxes of this set have a much higher volume than the box in the European industry norm. We used two different approaches to this generalization.

In a grid based approach, we used the same idea as above and discretized the search space by aligning the placements of the boxes to a grid. As we have boxes of different volumes, the independent set problem generalizes to a weighted version, which we can solve for rather coarse grids. As the lengths of the boxes are not multiples of our grid size, we have to round them. We get the best results, if we round the lengths down. Hence we compute an infeasible packing, but it usually can be made feasible by a physical simulation of the contacts between the boxes and the trunk.

We developed a further approach, which is not based on a grid. We start with a simplification of the volume of the trunk. For any type of a box and any of the axis-oriented orientations, we describe the set of feasible centers for this box in this orientation by a convex region that over-approximates this set and a set of convex forbidden regions, called obstacles, to correct this over-approximation. Then we use an algorithm that enumerates all combinations of boxes together with relative orientations between boxes and obstacles. We can test whether we can pack such a combination into our approximation of the trunk by solving a linear program. As we simplified the volume of the trunk, we can again find infeasible packings. Hence we legalize the packings by our physical simulation.

Both approaches normally find better packings as the manually created packings of the experts.

References

[1] R. Battiti and M. Protasi. Reactive local search for the maximum clique problem. *Algorithmica*, 29(4):610–637, 2001.

[2] K. Daniels and V. J. Milenkovic. Column-based strip packing using ordered and compliant containment. In *1st ACM Workshop on Applied Computational Geometry (WACG)*, 1996, pp. 33–38.

[3] Deutsches Institut für Normung e.V., ed. *DIN 70020, Teil 1, Straßenfahrzeuge; Kraftfahrzeugbau.* Deutsches Institut für Normung e.V., 1993. February 1993 revision.

[4] F. Eisenbrand, S. Funke, A. Karrenbauer, J. Reichel, and E. Schömer. Packing a trunk - now

with a twist! In S. N. Spencer, ed., *Proceedings SPM 2005 ACM Symposium on Solid and Physical Modeling*, Cambridge, USA, June 2005, pp. 197–206. ACM.

[5] F. Eisenbrand, S. Funke, J. Reichel, and E. Schömer. Packing a trunk. In G. Di Battista and U. Zwick, eds., *Algorithms - ESA 2003: 11th Annual European Symposium*, Budapest, Hungary, September 2003, *LNCS 2832*, pp. 618–629. Springer.

[6] R. F. Fowler, M. S. Paterson, and S. L. Tanimoto. Optimal packing and covering in the plane are NP-complete. *Information Processing Letters*, 12:133–137, 1981.

[7] V. J. Milenkovic. Rotational polygon containment and minimum enclosure using only robust 2d constructions. *Computational Geometry*, 13(1):3–19, 1999.

- [8] J. Reichel. *Combinatorial Approaches to Trunk Packing*. Phd thesis, Universität des Saarlandes, July 2006.
- [9] J. H. Rieskamp. Automation and optimization of Monte Carlo based trunk packing. Diploma thesis, Universität des Saarlandes, Saarbrücken, 2005.

[10] Society of Automotive Engineers, ed. *SAE J1100, Motor Vehicle Dimensions*. Society of Automotive Engineers, 2001. February 2001 revision.

[11] B. Verweij and K. Aardal. An optimization algorithm for maximum independent set with applications in map labelling. In *Proc. of the 7th Annual European Symposium (ESA '99), Prague, Czechia*, 1999, pp. 426–437.

Algorithms for longer OLED Lifetime

Investigators: Friedrich Eisenbrand and Andreas Karrenbauer

This research was partially developed in NWG2 and can be found in Section 19.3.1.

12.4.5 Algorithms for Bioinformatics

Bioinformatics is a relatively new and very important area of computer science. Many of the basic problems in this area are NP-hard combinatorial optimization problems. We give new exact algorithms for two of the most important problems in this area that outperform all former exact algorithms.

Computing Steiner Minimum Trees in Hamming Metric

Investigators: Ernst Althaus and Rouven Naujoks

A fundamental class of problems in Computational Biology is the reconstruction of phylogenetic trees which reads as follows: Given a set of species one wants to determine their ancestral relationship along a tree. In order to build such a tree we compare specific features of the species under the natural assumption that species with similar features are closely related. In modern phylogeny these features are defined by DNA or protein sequences.

The two probably most favorite and prominent versions are the *maximum likelihood* and the *maximum parsimony* methods. In each of them the goal is to find trees minimizing a certain cost function. In the maximum parsimony problem we want to find a tree that minimizes the number of changes of the genetic code along the edges of the tree. This can be mathematically restated as the problem of finding a *Steiner minimum tree* in *Hamming*

metric which is probably one of the most studied combinatorial optimization problems and which is known to be APX-hard. We have developed an algorithm (see [2]) for solving this problem exactly whose implementation clearly outperforms all existing solvers. Our approach is based on a branch-and-bound approach using a new way of tree enumeration combined with a variety of different pruning methods.

Multiple Sequence Alignment

Investigators: Ernst Althaus and Stefan Canzar

The comparison of a set of biological sequences (strings of DNA, RNA or amino acids) with the objective of detecting highly conserved subregions or common patterns is certainly one of the dominant problems in computational biology.

The *Multiple Sequence Alignment* (MSA) provides the mathematical framework in which those problems can be treated. An alignment of two strings s and t is obtained by inserting spaces into either string such that the two resulting sequences s' and t' are of the same length. Opposite letters from from s and t contribute a score that is specified by a pairwise scoring matrix. Maximal sequences of spaces in either string contribute a score depending on their length. The problem calls for an alignment whose overall score is maximized. In the multiple sequence alignment problem we insert spaces into the given strings such that all strings have the same length. The score of such an alignment is the sum all induced pairwise alignments.

In [1] Althaus et al. propose a branch-and-cut algorithm for the MSA problem based on an integer linear programming (ILP) formulation given by Reinert in [3]. As solving the LP-relaxation is by far the most expensive part of the algorithm and even not possible for moderately large instances, we propose a Lagrangian approach to approximate the linear program and utilize the resulting bounds on the optimal value in a branch-and-bound framework. Our experiments show that our implementation, although preliminary, outperforms all exact algorithms for the multiple sequence alignment problem. Furthermore, the quality of the alignments are among the best computed so far.

References

[1] E. Althaus, A. Caprara, H.-P. Lenhof, and K. Reinert. Multiple sequence alignment with arbitrary gap costs: Computing an optimal solution using polyhedral combinatorics. In T. Lengauer, ed., *Proceedings of the European Conference on Computational Biology (ECCB 2002)*, Saarbrücken, October 2002, *Bioinformatics*, vol. 18, pp. S4–S16. Oxford University Press.

- [2] 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.

[3] K. Reinert. *A polyhedral approach to sequence alignment problems.* Phd thesis, Universität des Saarlandes, August 1999.

12.5 Computational Geometry

Coordinators: Stefan Funke and Joachim Giesen

Our work in this area focuses both on the theoretical investigation of fundamental questions as well as on the application of known computational geometry techniques to other application domains. The efficient and exact implementation of geometric algorithms on curved objects is also one of the key research themes in this context but is described in more detail in the separate Section 12.8 on Software Systems. This section is organized in the following themes: sample-based geometry, applications in wireless networking, and fundamental algorithms and data structures.

Exemplary for the sample-based geometry theme is a result which establishes a relationship between the topology of a smooth surface and the structure of the neighborhood graph of a sufficiently dense point sample from the surface. A general theme in the area of wireless networking is the extraction of geometric or topological properties of the network deployment if no location information is available at the network nodes. As a fundamental result also relevant for the efficient and exact implementation of geometric algorithms on curved objects (also see Section 12.8), we show how to snap-round Bézier curves.

12.5.1 Sample-Based Geometry

A shape can often be well represented by a finite sample. For example it is possible to infer almost all topological and approximate geometric information about a smooth surface from a dense enough sample. In fact by today there are quite a few algorithms known that compute a surface (most often a piecewise linear surface) from a dense sample of a smooth surface that is homeomorphic to the original surface and shares many of its geometric properties. Similarly a dual structure of a surface, namely the medial axis (some sort of skeleton) of the volume bounded by a smooth surface can also to some extent be topologically correctly reconstructed from a finite sample of the surface. Here topological correctness refers to the weaker concept of homotopy equivalence compared to being homomorphic as it is usually the goal in surface reconstruction.

Cycle Bases of Graphs and Sampled Manifolds

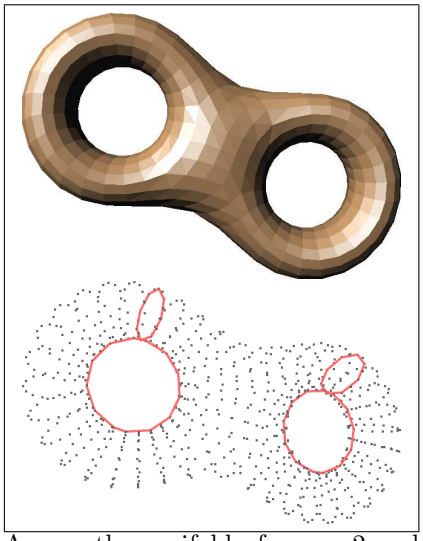
Investigators: Kanela Kaligosi, Kurt Mehlhorn, Dimitrios Michail, and Evangelia Pyrga in collaboration with Craig Gotsman

Point samples of a surface in $\mathbb{R}^3$ are the dominant output of a multitude of 3D scanning devices. The usefulness of these devices rests on being able to extract properties of the surface from the sample. This work [2] shows that, under certain sampling conditions, the minimum cycle basis of a nearest neighbor graph of the sample encodes topological information about the surface and yields bases for the trivial and non-trivial loops of the surface. Consider a compact manifold S in $\mathbb{R}^3$ and a finite set of points P in S. A common approach into treating such point samples is to first construct some neighboring graph.

One such popular graph is the k-nearest neighbor graph; an undirected graph with vertex set P and an edge between two sample points a and b if b is one of the k points closest to a and a is one of the k points closest to b.

We study the minimum cycle basis of the k-nearest neighbor graph when S is a compact smooth manifold and P is a sufficiently dense sample.

We show that for suitably nice samples of smooth manifolds of genus g and sufficiently large k, the k-nearest neighbor graph G_k has a cycle basis consisting only of short (= length at most $2(k+3)$) and long (= length at least $4(k+3)$) cycles. Moreover, the minimum cycle basis is such a basis and contains exactly $m-(n-1)-2g$ short cycles and $2g$ long cycles. The short cycles span the subspace of trivial loops and the long cycles form a homology basis.



A smooth manifold of genus 2 and the 4 largest cycles of a minimum cycle basis of the 8-nearest neighbor graph.

Thus, the MCB of G_k reveals the genus of S and also provides a basis for the set of trivial cycles and a set of generators for the non-trivial cycles of S. We also validate our results with experiments.

Medial Axis Approximation from Inner Voronoi Balls: A Demo of the Mesecina Tool

Investigators: Joachim Giesen in collaboration with Balint Miklos and Mark Pauly

We used Mesecina to explore the medial axis of shapes with smooth boundaries in the plane. The idea – as programmatic in sampled based geometric modeling – is to sample the boundary of the shape and compute a medial axis approximation from the sample only. The approximation builds on computing a set of balls that approximates the shape quite well and then compute the medial axis of the union of these balls. In contrast to the standard pipeline in our special setting the medial axis of the union of balls can be computed from the Voronoi diagram of sample points (provided the sampling is sufficiently dense).

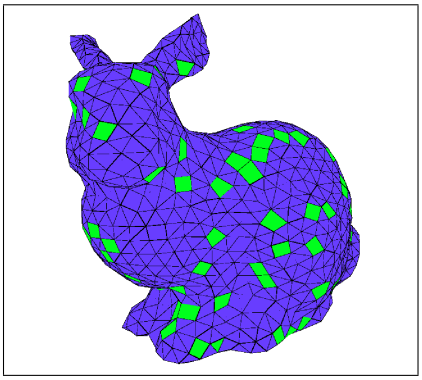
We used Mesecina to create the video demo, which shows a standard and more general approach to compute the medial axis of a union of balls and our approach which is simpler, faster and numerically more stable (but only works for densely sampled shapes with smooth boundary). See [3].

Ignoring Slivers – in Practice and with a Guarantee

Investigators: Daniel Dumitriu, Stefan Funke, Martin Kutz, and Nikola Milosavljevic

Near-degenerate configurations of sample points called *slivers* still pose a challenge to many current surface reconstruction algorithms. Furthermore, the geometric predicates required for most surface reconstruction algorithms are computationally rather demanding, in particular if exactness of the outcome is to be guaranteed. In this work we propose a new approach for reconstructing a 2-manifold from a point sample in $\mathbb{R}^3$ that addresses both these issues.

Compared to previous algorithms our approach is novel in that it throws away geometry information early on in the reconstruction process and mainly operates combinatorially on a graph structure (hence reducing the cost of evaluating complicated predicates). Furthermore it is very conservative in creating adjacencies between samples in the vicinity of slivers, still we can prove that the resulting reconstruction faithfully resembles the original 2-manifold. While the theoretical proof requires an extremely high sampling density our prototype implementation of the approach produces surprisingly good results on typical sample sets and seems to have great potential in particular in parallel computing or external memory scenarios. This is work currently under progress, [1].



Conservative adjacency creation leads to non-triangular faces.

References

- [1] D. Dumitriu, S. Funke, M. Kutz, and N. Milosavljevic. Avoiding slivers - in practice and with a guarantee. 2007.
- [2] C. Gotsman, K. Kaligosi, K. Mehlhorn, D. Michail, and E. Pyrga. Cycle bases of graphs and sampled manifolds. *Computer Aided Geometric Design*, 2007. To appear.
- [3] 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 2007)*, Gyeongju, South Korea, 2007. ACM. To appear.

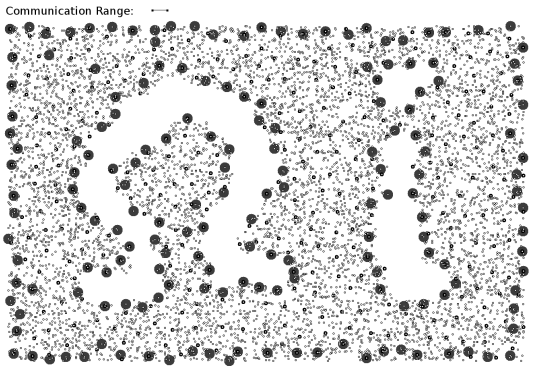
12.5.2 Applications in Wireless Communication

In recent years wireless network technology has gained tremendous importance. While the spatial aspect was already of interest in the wired network world due to cable costs etc., it has far more influence on the design and operation of wireless networks. Whether or not a node can communicate with another node is strongly correlated with the Euclidean distance between the two nodes. Hence problems in this area are prime candidates for the use of techniques from computational geometry.

Hole-Detection in Wireless Networks

Investigators: Stefan Funke and Christian Klein

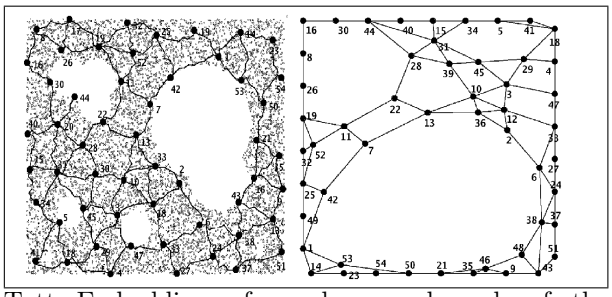
Wireless sensor networks typically consist of small, very simple network nodes without any positioning device like GPS. After an initialization phase, the nodes know with whom they can talk directly, but have no idea about their relative geographic locations. We examine how much geometry information is nevertheless hidden in the communication graph of the network: Assuming that the connectivity is determined by the well-known unit-disk graph model, we prove that using an extremely simple linear-time distributed algorithm one can identify nodes on the boundaries of holes of the network. Hence, there is enough information hidden in the communication graph to identify holes in the network. We have theoretically



shown that our algorithm produces a correct result under certain assumptions on the density of the network nodes. Our actual implementation shows that the algorithm works well under less stringent conditions, too. This result was published in [2].

Network Sketching

Investigators: Stefan Funke and Nikola Milosavljevic



Tutte-Embedding of a planar subgraph of the communication graph.

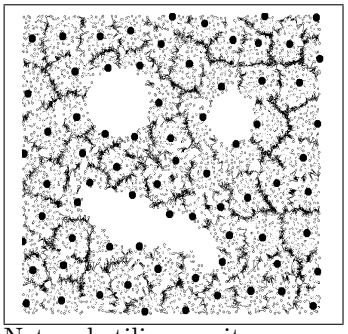
Extending the work in [2] we show that for a communication graph determined by the well-known unit-disk graph model a very simple distributed algorithm can identify a large, provably planar subgraph of the communication graph that faithfully reflects the topology of the network. This planar subgraph can then be embedded using a simple distributed rubber-banding procedure, finally obtaining *virtual coordinates* for the nodes of the subgraph which can be instrumented for various protocols based on geographic location information. That is, there is enough geometry information hidden in the connectivity structure not only to identify topological features like network holes (as it was done in [2]) but even enough information to compute a *sketch* of the network layout. Our simulation results indicate that the algorithm works very well even for very sparse network deployments and produces network sketches that come close to the original layout. This result was published in [5].

Guaranteed-delivery Geographic Routing under imprecise Node Locations

Investigators: Stefan Funke and Nikola Milosavljevic

Geographic routing protocols like GOAFR or GPSR rely on *exact* location information at the nodes, as when the greedy routing phase gets stuck at a local minimum, they require, as a fallback, a planar subgraph whose identification, in all existing methods, depends on exact node positions. In practice, however, location information at the network nodes is hardly precise;

be it because the employed location hardware, such as GPS, exhibits an inherent measurement imprecision, or because the localization protocols which estimate positions of the network nodes cannot do so without errors. In this paper we propose a novel naming and routing scheme that can handle the uncertainty in location information. It is based on a macroscopic variant of geographic greedy routing, as well as a macroscopic planarization of the communication graph. If an upper bound on the deviation from true node locations is available, our routing protocol guarantees delivery of messages. Due to its macroscopic view, our routing scheme also produces shorter and more load-balanced paths than common geographic routing schemes, in particular in sparsely connected networks or in the presence of obstacles, see [4].



Network tiling as it appears in the construction of our routing scheme.

Distance-sensitive Information Brokerage in Sensor Networks

Investigators: Stefan Funke in collaboration with Leonidas Guibas, An Nguyen, and Yusu Wang

In a sensor network information from multiple nodes must usually be aggregated in order to accomplish a certain task. A natural way to view this information gathering is in terms of interactions between nodes that are *producers* of information, e.g., those that have collected data, detected events, etc., and nodes that are *consumers* of information, i.e., nodes that seek data or events of certain types. In [1] we propose efficient schemes allowing consumer and producer nodes to discover each other so that the desired information can be delivered quickly to those who seek it. Here, efficiency means both limiting the redundancy of where producer information is stored, as well bounding the consumer query times. We introduce the notion of *distance-sensitive information brokerage* and provide schemes for efficiently bringing together information producers and consumers at a cost proportional to the separation between them – even though neither the consumers nor the producers know about each other beforehand.

Applications of Compact Network Synopses

Investigators: Stefan Funke, Sören Laue, and Rouven Naujoks

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.

Several papers have treated theses problems before and derived approximation algorithms for them as these problems are in general NP-hard. However, the running time of these algorithms is still quite high, in some cases even triply exponential in the number of radio nodes. As the size of a wireless network becomes larger and larger, previous algorithms are not able anymore to deal with networks of this size. For this reason, we designed a new

concept – the *network synopsis*. A network synopsis in principle is a sketch of the network which is a good representation of the total network. Thus, the problem only needs to be solved on a much smaller instance which in turn gives a good solution ($(1+\epsilon)$-approximation) for the original network instance.

Among the wide range of power assignment problems we considered was the k-hop broadcasting problem, where one wants to broadcast a message from a specific source node to all other nodes within at most k hops. For this problem we were able to construct a constant size network synopsis [3]. This allows us to solve the problem in time that is linear in the network size which is a drastic improvement upon previous results. Another problem for which we were able to show the existence of a constant size network synopsis is the k-set broadcast problem. Here one wants to broadcast a message form a given source node to all other nodes and at most k radio nodes are allowed to send. The reason for keeping the number of sending radio nodes small is the reduction of the probability of a network failure. Another problem we considered was the problem of assigning transmission powers to k radio stations such that each input node is within reach of at least one of them. Again, we were able to construct a constant size network synopsis and hence find an $(1+\epsilon)$-approximate solution in linear time.

We believe that the concept of a network synopsis can be generalized to a wide range of other network problems and will be then useful to speed up existing algorithms.

References

- [1] S. Funke, L. Guibas, A. Nguyen, and Y. Wang. Distance-sensitive information brokerage in sensor networks. In P. B. Gibbons, T. F. Abdelzaher, J. Aspnes, and R. Rao, eds., *Distributed Computing in Sensor Systems, Second IEEE International Conference, DCOSS 2006*, San Francisco, USA, 2006, *LNCS 4026*, pp. 234–251. Springer.
- [2] S. Funke and C. Klein. Hole detection or: "how much geometry hides in connectivity?". In N. Amenta and O. Cheong, eds., *Proceedings of the 22nd Annual Symposium on Computational Geometry, SCG'06*, Sedona, Arizona, USA, 2006, pp. 377–385. ACM.
- [3] S. Funke and S. Laue. Bounded-hop energy-efficient broadcast in low-dimensional metrics via coresets. In W. Thomas and P. Weil, eds., *24th Annual Symposium on Theoretical Aspects of Computer Science (STACS 2007)*, Aachen, Germany, 2007, *LNCS 4393*, pp. 272–283. Springer.
- [4] S. Funke and N. Milosavljevic. Guaranteed-delivery geographic routing under uncertain node locations. In *11th IEEE Conference on Computer Communication (INFOCOM)*, Anchorage, USA, 2007. IEEE. to appear.
- [5] S. Funke and N. Milosavljevic. Network sketching or: "how much geometry hides in connectivity? - part ii". In *Proceedings of the eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-07)*, New Orleans, USA, January 2007, pp. 958–967. SIAM.

Conflict-free Coloring of Rectangles

Investigators: Deepak Ajwani, Khaled Elbassioni, Sathish Govindarajan, and Nabil Mustafa in collaboration with Saurabh Ray

The study of conflict-free coloring is motivated by the frequency assignment problem in wireless networks. A wireless network is a heterogeneous network consisting of *base stations*

and *agents*. The base stations have a fixed location, and are interlinked via a fixed backbone network, while the agents are typically mobile and can connect to the base stations via radio links. The base stations are assigned fixed frequencies to enable links to agents. The agents can connect to any base station, provided that the radio link to that particular station has good reception. Good reception is only possible if i) the base station is located within range, and ii) no other base station within range of the agent has the same frequency assignment (to avoid interference). Thus the fundamental problem of frequency-assignment in cellular networks is to assign frequencies to base stations, such that an agent can always find a base station with unique frequency among the base stations in its range. Naturally, due to cost, flexibility and other restrictions, one would like to minimize the total number of assigned frequencies.

The problem is mathematically described as follows. Let $P \subseteq \mathbb{R}^2$ be a set of points and $\mathbb{R}$ be a set of rectangular ranges. A *conflict-free* coloring (CF-coloring in short) of P w.r.t. the range $\mathbb{R}$ is an assignment of a color to each point $p \in P$ such that for any range $T \in \mathbb{R}$ with $T \cap P \neq 0$, the set $T \cap P$ contains a point of unique color. Naturally, the goal is to assign a conflict-free coloring to the points of P with the *smallest* number of colors possible.

Har Peled et al [5] gave a simple algorithm that uses $O(\sqrt{n})$ colors for CF-coloring rectangular ranges. This can be further improved to $O(\sqrt{n \log\log n / \log n})$ using the sparse neighborhood property of the conflict-free graph, as independently observed in [2, 6]. Recent works [3, 4] have shown that one can obtain better upper bounds for special cases of this problem.

We [1] present a simple algorithm that CF-colors P with respect to rectangle ranges $\mathbb{R}$, using $\tilde{O}(n^{.382+\epsilon})$ colors, in expected polynomial time, for any arbitrarily small $\epsilon > 0$. Our main tool for proving this theorem is a probabilistic coloring technique, introduced in [4], that can be used to get a coloring with weaker properties, which we call *quasi-conflict-free* coloring. This is combined with dominating sets, monotone sequences, and careful griding of the point set, in a recursive way, to obtain our result.

References

- [1] D. Ajwani, K. 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, June 2007, Proceedings. ACM. To Appear.

  [2] N. Alon, M. Krivelevich, and B. Sudakov. Coloring graphs with sparse neighborhoods. *J. Combinatorial Theory Ser. B.*, 77:73–82, 1999.

  [3] K. Chen. How to play a coloring game against a color-blind adversary. In *SCG '06: Proceedings of the twenty-second Annual Symposium on Computational Geometry*, New York, NY, USA, 2006, pp. 44–51. ACM Press.

- [4] K. Elbassioni and N. H. Mustafa. Conflict-free colorings of rectangle ranges. In B. Durand and W. Thomas, eds., *STACS 2006, 23rd Annual Symposium on Theoretical Aspects of Computer Science*, Marseille, France, February 2006, *LNCS 3884*, pp. 254–263. Springer.

  [5] S. Har-Peled and S. Smorodinsky. Conflict-free coloring of points and simple regions in the plane. *Discrete & Comput. Geom.*, 34:47–70, 2005.

  [6] J. Pach and G. Toth. Conflict free colorings. In *Discrete & Comput. Geom., The Goodman-Pollack Festschrift*. Springer Verlag, Heidelberg, 2003.

12.5.3 Algorithms

Computational geometry is an extremely broad field which deals with problems in combinatorial discrete geometry, polyhedral combinatorics, geometric modelling or such classical questions like the Euclidean travelling salesman problem. In the following subsection we give a brief overview of our results in the field which do not fall under the general headings of Sample-based Geometry (12.5.1) or Applications in Wireless Communication (12.5.2). Amongst other results we provide a generalization of the centerpoint theorem, show how to approximate the Euclidean TSP with neighborhoods, and investigate the structure of certain polyhedra.

An Optimal Generalization of the Centerpoint Theorem.

Investigators: Nabil Mustafa and Saurabh Ray

The centerpoint theorem is one of the fundamental combinatorial results in discrete geometry, with uses in geometric algorithms [3, 5], large-scale computing [2] and several other areas. It states the following: Given a set P of n points in the plane, there exists a point c such that any convex object containing more than $2n/3$ points of P contains c. Furthermore, this bound is tight.

In [4] we look at generalization of the above statement to larger number of points. For example, is it possible to find two points c_1 and c_2 in the plane such that any convex object containing at least $n/2$ points must contain either c_1 or c_2? This problem had been studied by Aronov *et al.* [1], who proved that if each convex object contains at least $5n/8$ points of P, then they can be hit by two points. They also provide a lower-bound of $5n/9$.

We present a general procedure that gives the following results: one can hit all convex objects containing more than $4n/7$ points with 2 points. Furthermore, we prove that this bound is tight. Similar results are derived for larger number of points. In particular, we show that if each convex object contains more than $20n/41$ points, then five points suffice. This improves a natural way of adding five points [1] which gives the worse $n/2$-bound: find two lines (using ham-sandwich cut) which partition the point set into four regions with $n/4$ points in each. Add the intersection point of the lines along with the centerpoints of the four regions.

Besides giving an optimal generalization of the centerpoint theorem, our method in fact also gives an elementary short new proof of the centerpoint theorem (and also Helly's theorem).

References

[1] B. Aronov, F. Aurenhammer, F. Hurtado, S. Langerman, D. Rappaport, S. Smorodinsky, and C. Seara. Small weak epsilon nets. In *Proc. 17th Canadian Conference on Computational Geometry*, 2005.

[2] G. Miller, S. Teng, W. Thurston, and S. Vavasis. Automatic mesh partitioning. *Workshop on Sparse Matrix Computations: Graph Theory Issues and Algorithms*, 1993.

[3] G. L. Miller, S.-H. Teng, W. P. Thurston, and S. A. Vavasis. Separators for sphere-packings and nearest neighbor graphs. *Journal of ACM*, 44:1–29, 1997.

• [4] N. H. Mustafa and S. Ray. An optimal generalization of the centerpoint theorem and its extensions. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry (SCG 2007)*, Gyeongju, South Korea, 2007. ACM.

[5] F. F. Yao. A 3-space partition and its applications. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, 1983, pp. 258–263.

Construction of Weak ϵ-nets with Small Basis

Investigators: Nabil Mustafa and Saurabh Ray

Given a set system $(X, \mathcal{F})$, where X is the base set, and $\mathcal{F}$ is a family of subsets of X, the general ϵ-net problem asks for a small subset X' of X such that for every set $S \in \mathcal{F}$ containing at least $\epsilon|X|$ elements, $X' \cap S \neq \emptyset$. In a celebrated result, Haussler and Welzl [1] showed that if the set system has finite VC-dimension, then picking a random sample from X of size $O(1/\epsilon \log 1/\epsilon)$ (constant dependent linearly on the VC-dimension of the set system) yields an ϵ-net with some constant probability.

Unfortunately, the existence of small ϵ-nets is no longer true for set systems of infinite VC-dimension. For example, it is easy to see that any ϵ-net with respect to convex ranges must have at least $(1-\epsilon)n$ points of P if P is in convex position. The concept of *weak* ϵ-nets with respect to *convex ranges* was introduced by Haussler and Welzl [1] in their seminal paper: the restriction that the points of ϵ-net be a subset of X is dropped. Weak ϵ-nets (w.r.t. convex ranges) have found several applications in discrete and combinatorial geometry [2].

The long-standing open problem in this area has been to show the existence of weak ϵ-nets in $\Re^d$ with size $o(1/\epsilon^d)$. In fact, the current conjecture [3] is that optimal weak ϵ-nets should have size $O(1/\epsilon\ \log(1/\epsilon))$ (hidden constants depend on the dimension) in $\Re^d$.

In [4], we first observe that a weak ϵ-net of P of size k is completely described by $O(d^2k)$ points of P. Hence if weak ϵ-nets do have size $O(1/\epsilon)$ in any dimension, then there must exist $O(1/\epsilon)$ (hidden constants depend on d) points of P from which it is constructed. So a possible first step towards solving the conjecture is to show this linear dependence on points of P. Unfortunately all known constructions of weak ϵ-nets use $\Omega(1/\epsilon^d)$ input points.

In [4], we answer the above question in the affirmative, showing that for every point set P, there exists a set of $O(1/\epsilon \log 1/\epsilon)$ points in $\Re^d$ from which one can construct a weak ϵ-net for P. So while the size of weak ϵ-nets we compute involve d, their description (i.e., points used to construct them) is in fact near-linear in $1/\epsilon$. The proof establishes the following interesting relation between strong ϵ-nets and weak ϵ-nets: In $\Re^2$, take an ϵ-net with respect to the intersection of every six halfplanes. Then *only* from these $O(1/\epsilon \log 1/\epsilon)$ points, one can construct a weak ϵ-net of size $O(1/\epsilon^3 \log^3 1/\epsilon)$. Similarly, we show that by random sampling $O(1/\epsilon \log 1/\epsilon)$ points in $\Re^3$, and taking some function of these sampled points, one gets a weak ϵ-net of size $O(1/\epsilon^5 \log 1/\epsilon^5)$. For P in $\Re^d$, take a random sample of size $O(1/\epsilon \log 1/\epsilon)$ (with only the constant depending on d). Then another product function of these sampled points yields an ϵ-net with size $O(1/\epsilon^{d^2})$.

References

[1] D. Haussler and E. Welzl. Epsilon-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151, 1987.

[2] J. Matousek. *Lectures in Discrete Geometry.* Springer-Verlag, New York, NY, 2002.

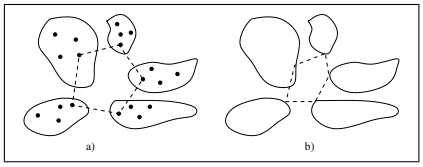
[3] J. Matousek and U. Wagner. New constructions of weak epsilon-nets. *Discrete & Computational Geometry*, 32(2):195–206, 2004.

- [4] N. H. Mustafa and S. Ray. 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 2007)*, Gyeongju, South Korea, 2007. ACM.

Approximating Euclidean TSP with Neighborhoods

Investigators: Khaled Elbassioni, Nabil H. Mustafa, Aleksei V. Fishkin, and René Sitters

In the Euclidean group *Traveling Salesman Problem* (TSP), we are given a set of points P in the plane and a set of n connected regions (neighborhoods), each containing at least one point of P. The objective is to find a tour of minimum length that visits at least one point in each region. This generalizes both the classical Euclidean TSP and the group Steiner tree problems with applications in VLSI-design, and other routing-related applications. The difficulty of the problem varies depending on whether the given regions are disjoint/non-disjoint, fat/non-fat, have comparable/different diameters, etc. One can also distinguish between two cases: the one in which $P = \mathbb{R}^2$, i.e. the TSP tour can hit any point of the connected region, referred to as the *continuous* case (or the TSP with neighborhoods), and the one in which the tour is allowed to hit the region only in one of the specified points (i.e. P is finite)), referred to as the *discrete* case. In [1, 2], we consider several interesting variants of the problem, and give approximation algorithms of different guarantees depending on the variant considered. In particular, we give (i) an $O(\alpha)$-approximation algorithm for the case when the regions are disjoint α-fat objects with possibly varying size; (ii) an $O(\alpha^3)$-approximation algorithm for intersecting α-fat regions with comparable diameters. These results also apply to the continuous case, in which the sought TSP tour can hit each region at any point. (iii) For the general continuous case of TSP with connected neighborhoods, we give a simple $O(\log n)$-approximation algorithm. The most distinguishing features of these algorithms are their simplicity and low running-time complexities.



a) a tour in Euclidean Group TSP b) a tour in Euclidean TSP with Neighborhoods.

References

- [1] K. Elbassioni, A. Fishkin, N. H. Mustafa, and R. Sitters. Approximation algorithms for euclidean group tsp. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, eds., *Automata, languages and programming, 32nd International Colloquim, ICALP 2005*, Lisbon, Portugal, 2005, *LNCS 3580*, pp. 1115–1126. Springer.

- [2] K. Elbassioni, A. V. Fishkin, and R. Sitters. On approximating the TSP with intersecting neighborhoods. In T. Asano, ed., *Algorithms and Computation, 17th International Symposium, ISAAC 2006*, Kolkata, India, 2006, *LNCS 4288*, pp. 213–222. Springer.

Vertex Enumeration for Polyhedra

Investigators: Khaled Elbassioni in collaboration with Leonid Khachiyan, Endre Boros, Konrad Borys, Vladimir Gurvich, and Kazuhisa Makino

An important and a longstanding open question in linear programming is whether there exists an efficient way to find k different vertices of a given polyhedron $P = \{x \in \mathbb{R}^n \mid Ax \leq b\}$, in time polynomial in the size of A, b, and k. In this work, we made a partial progress towards answering this question, by proving that the problem is NP-hard in the case of unbounded polyhedra [2]. In other words, we show that there exists no algorithm which can generate k vertices of P, in time polynomial in the size of A, b and k, unless P=NP. This is obtained by showing that generating all negative cycles of a given weighted directed graph is an NP-hard enumeration problem. The corresponding problem for polytopes (bounded polyhedra) remains open.

In [1], we give evidence that the method used to list the vertices of 0/1-polytopes (i.e those in which every vertex has only binary components) in polynomial time is unlikely to extend to 0/1 polyhedra. This is obtained by showing that the basic *extension* subproblem on which the efficiency of the method relies, is generally NP-hard. The reduction uses polyhedra associated with network matrices A, with the right hand side b being the vector of all ones. On the other hand, we show (by using another method) that the vertices of such polyhedra, or their duals, can be enumerated in incremental polynomial time using polynomial space.

References

- [1] E. Boros, K. Elbassioni, V. Gurvich, and K. Makino. Generating vertices of polyhedra and related monotone generation problems. February 2007.
- [2] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, and V. Gurvich. Generating all vertices of a polyhedron is hard. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'06*, Miami, FL, USA, January 2006, pp. 758–765. ACM / SIAM.

Upper Bound on the Number of Vertices of Polyhedra with $0,1$-Constraint Matrices

Khaled Elbassioni, Zvi Lotker, Raimund Seidel

In [1], we give upper bounds for the number of vertices of the polyhedron $P(A,b) = \{x \in \mathbb{R}^d : Ax \leq b\}$ when the $m \times d$ constraint matrix A is subjected to certain restriction. For instance, if A is a 0/1-matrix, then there can be at most $d!$ vertices and this bound is tight, or if the entries of A are non-negative integers so that each row sums to at most C, then there can be at most C^d vertices. These bounds are consequences of a more general theorem that the number of vertices of $P(A,b)$ is at most $d! \cdot W/D$, where W is the volume of the convex hull of the zero vector and the row vectors of A, and D is the smallest absolute value of any non-zero $d \times d$ subdeterminant of A.

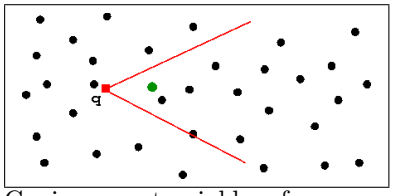
References

- [1] K. Elbassioni, Z. Lotker, and R. Seidel. Upper bound on the number of vertices of polyhedra with 0, 1-constraint matrices. *Information Processing Letters*, 100(2):69 – 71, October 2006.

(Approximate) Conic Nearest Neighbors and their induced Voronoi Diagram

Investigators: Stefan Funke, Domagoj Matijevic, Theocharis Malamatos, and Nicola Wolpert

In [1], we consider the following variant of the nearest neighbor search problem in Euclidean space: given a set of points S (and a cone C) we want to preprocess S such that for any query point $q \in \mathbb{R}^d$ we can determine an (approximate) nearest neighbor $s_q \in S$ that is contained in the cone C with apex at q. We examine the structure of the exact conic Voronoi diagram induced by the notion of conic proximity in 2 and 3 dimensions. Furthermore, we develop a data structure such that for an *arbitrary* simplicial cone *exact* cone queries can be reported in sub-linear time. As our main result, we show how to construct an approximate conic Voronoi diagram of size $O((n/\epsilon^d)\log(1/\epsilon))$ which allows for queries in time $O(\log(n/\epsilon))$.



Conic nearest neighbor for q.

References

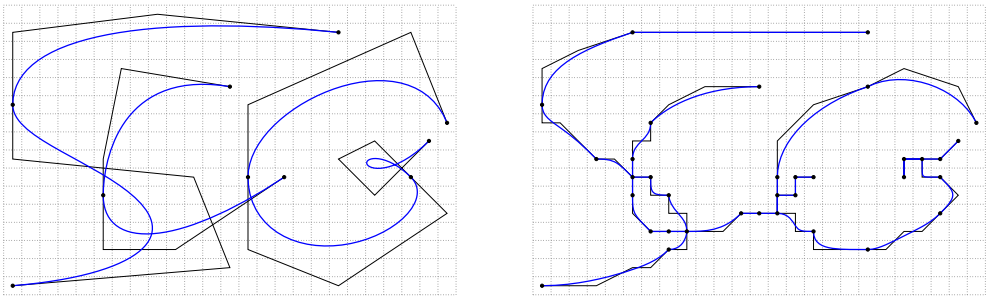
- [1] S. Funke, T. Malamatos, D. Matijevic, and N. Wolpert. (approximate) conic nearest neighbors and the induced voronoi diagram. In *18th Canadian Conference on Computational Geometry*, Kingston, Canada, 2006, pp. 23–26. School of Computing, Queen's University.

Snap Rounding of Bézier Curves

Investigators: Arno Eigenwillig, Lutz Kettner, and Nicola Wolpert

We propose a method [1] [2] to compute a geometric rounding of the planar arrangement induced by a set of Bézier curves. *Bézier curves* are polynomially parameterized curves $\mathbf{b}(t) = \sum_{i=0}^{n} \mathbf{b}_i B_i^n(t)$ $(0 \le t \le 1)$ that are expressed in terms of a geometrically meaningful control polygon $(\mathbf{b}_0, \ldots, \mathbf{b}_n)$ and Bernstein polynomials $B_i^n(t) = \binom{n}{i} t^i (1-t)^{n-i}$. Bézier curves are ubiquitous in geometric modelling and CAGD. A *geometric rounding* of an arrangement is a modification of the arrangement in which all numerical data (vertex and control point coordinates) are rounded, but only after the combinatorial structure has been adapted to guarantee that rounding does not invert topology; that is, small features may collapse, but no vertex crosses over an edge, etc.

Previously, geometric rounding has only been investigated for arrangements of straight-line objects. One particularly popular method for straight-line segments in the plane is *snap rounding*, see [3]. We have extended the techniques of snap rounding to control polygons of Bézier curves. We have then implanted them as a sound stopping criterion in the traditional method of intersection by repeated subdivision "up to some precision ε": subdivision stops if intersection points have been localized with sufficient precision for rounding and if the properties of straight-line snap rounding guarantee that rounding does not invert topology.



Seven input curves (left) and their snap-rounded arrangement (right).

Notice that our method computes the snap-rounded arrangement right away; we bypass the costly computation of the exact arrangement, which would involve irrational algebraic point coordinates. Nevertheless, we are able to give a topological guarantee on the output in terms of the true arrangement of the given inputs. This is akin to a bound on "forward error" in numerical analysis and thus rather different from the guarantee given by the Controlled Perturbation approach (cf. section 12.8.3), which is essentially a bound on "backward error" (i.e., on the perturbation of the inputs).

References

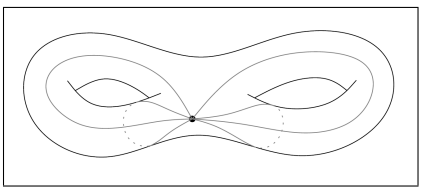
- [1] A. Eigenwillig, L. Kettner, and N. Wolpert. Snap rounding of Bézier curves. Research Report MPI-I-2006-1-005, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, December 2006.

- [2] A. Eigenwillig, L. Kettner, and N. Wolpert. Snap rounding of Bézier curves. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry (SCG 2007)*, Gyeongju, South Korea, 2007. ACM. To appear.

[3] L. J. Guibas and D. H. Marimont. Rounding arrangements dynamically. *Internat. J. of Comput. Geometry & Applications*, 8:157–176, 1998.

Shortest Non-Trivial Cycles on Bounded-Genus Surfaces

Investigator: Martin Kutz



System of loops on a 2-hole torus.

A combinatorial surface is a 2-dimensional manifold $\mathcal{M}$ together with a graph G embedded on $\mathcal{M}$ such that every face is a topological disk. The decomposition of such surfaces has recently gained growing attention in the field of computational topology. A central open question that arose in this context is: How to compute a shortest non-trivial cycle on a topological surface efficiently? Depending on the desired decomposition the question translates into the search for a non-contractible or a non-separating cycle. In [1] we present an almost linear time algorithm that solves both questions for orientable surfaces. The algorithm uses universal-cover constructions and makes extensive use of existing tools from the field.

References

- [1] M. Kutz. Computing shortest non-trivial cycles on orientable surfaces of bounded genus in almost linear time. In N. Amenta and O. Cheong, eds., *Proceedings of the 22nd Annual Symposium on Computational Geometry, SCG'06*, Sedona, Arizona, USA, 2006, pp. 430–437. ACM.

Geometric Minimum-Dilation Graphs

Investigator: Martin Kutz in collaboration with Rolf Klein

In [1] and [2] we consider a geometric graph G, drawn with straight lines in the plane. For every pair a, b of vertices of G, we compare the shortest-path distance between a and b in G (with Euclidean edge lengths) to their actual Euclidean distance in the plane. The worst-case ratio of these two values, for all pairs of vertices, is called the vertex-to-vertex dilation of G.

In [1] we prove that computing a minimum-dilation graph that connects a given n-point set in the plane, using not more than a given number m of edges, is an NP-hard problem, no matter if edge crossings are allowed or forbidden. In addition, we show that the minimum dilation tree over a given point set may in fact contain edge crossings.

All finite graphs of vertex-to-vertex dilation 1 have been classified and are closely related to the following iterative procedure. For a given point set $P \subset \mathbb{R}^2$, we connect every pair of points in P by a line segment and add all intersection points of these lines to P. Repeating this process infinitely often we obtain a limit point set $P^\infty \supseteq P$. P^∞ is finite if and only if P is contained in a vertex set of dilation 1. In [2] we show that for any finite point set P for which P^∞ is infinite, there exists a threshold $\lambda > 1$ such that P is not contained in in the vertex set of any finite plane graph of dilation at most λ.

References

- [1] 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.
- [2] R. Klein and M. Kutz. The density of iterated crossing points and a gap result for triangulations of finite point sets. In N. Amenta and O. Cheong, eds., *Proceedings of the 22nd Annual Symposium on Computational Geometry (SCG06)*, Sedona, Arizona, USA, 2007, pp. 264–272. ACM.

12.6 Algorithms for Advanced Models of Computation

Coordinator: Uli Meyer

Basic algorithmic research traditionally assumed some variant of the von Neumann model of computation with a single processor and uniform memory. However, more advanced models are by now an important established part of algorithmic research because many of the challenges in modern computer science have to do with communication, parallel and distributed computing, and memory hierarchies.

Topics covered in the last two years were external-memory algorithms (Section 12.6.1), streaming algorithms (Section 12.6.2), and approaches for lock-free memory allocation (Section 12.6.3). Note that algorithms for wireless networks (previously discussed under advanced models) are now covered in the geometry chapter (Section 12.5).

12.6.1 External-Memory Algorithms

Investigators: Deepak Ajwani, Andreas Beckmann, Sathish Govindarajan, Lutz Kettner, Uli Meyer and Vitaly Osipov in collaboration with Roman Dementiev, Debora Donato, Jörg Keller, Luigi Laura, Tamas Lukovszki, Anil Maheshwari, Stefano Millozzi, Peter Sanders, and Norbert Zeh

In the last two years with obtained improved results for breadth first search, depth first search, single-source shortest paths, and well-separated pair decompositions.

Breadth First Search

Despite the existence of simple linear time algorithms in the RAM model, BFS was considered non-viable for massive graphs because of the I/O cost it incurs. Since the RAM model does not capture the I/O costs, we consider the commonly accepted external memory model by Aggarwal and Vitter [1].

It assumes a two level memory hierarchy with a fast internal memory and a slow external memory. We define M ($< n + m$) to be the number of vertices/edges that fit into internal memory, and B to be the number of vertices/edges that fit into a disk block. In an I/O operation, one block of data is transferred between disk and internal memory (Figure 12.8). The measure of performance of an algorithm is the number of I/Os it performs. The number of I/Os needed to read N contiguous items from disk is $\text{scan}(N) = \Theta(N/B)$. The number of I/Os required to sort N items is $\text{sort}(N) = \Theta((N/B)\log_{M/B}(N/B))$. For all realistic values of N, B, and M, $\text{scan}(N) < \text{sort}(N) \ll N$.

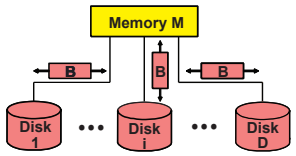


Figure 12.8: I/O Model

Munagala and Ranade [12] and later Mehlhorn and Meyer [9] gave efficient algorithms (referred to as MR_BFS and MM_BFS, respectively) for computing BFS level decompositions in an external memory model. MR_BFS takes $O(n+\text{sort}(n+m))$ I/Os, while MM_BFS takes $O(\sqrt{\frac{n(n+m)}{B}} + \text{sort}(n+m)\log\log\frac{B\cdot n}{m})$. In [2], we presented our implementation of MR_BFS and the randomized variant of MM_BFS using the external memory library STXXL [6] and gave a comparative study of the two algorithms on various graph classes. We demonstrated that the usage of these algorithms along with disk parallelism and pipelining can alleviate the I/O bottleneck of BFS on many small diameter graph classes, thereby making the BFS viable for these graphs. As a real world example, the BFS level decomposition of an external web-crawl based graph of around 130 million nodes and 1.4 billion edges was computed in less than 4 hours using a single disk and 2.3 hours using four disks.

However, both MR_BFS and the randomized variant of MM_BFS take *days* on large diameter graphs. In [3], we show that the deterministic variant of MM_BFS coupled with a heuristic can be used for computing the BFS level decomposition of even large diameter graphs in a few *hours*. MM_BFS decomposes the graph into low diameter clusters and maintains an efficiently accessible pool of adjacency lists required in the next few levels. For many large diameter graphs, the pool fits into the internal memory most of the time. By keeping the portion of the pool that fits into the internal memory as a multi-map hash table and by using caching of adjacency lists with two level hierarchy of clusters, we significantly improve

the performance of the external BFS algorithm while keeping the worst case I/O bounds of MM_BFS.

Single-Source Shortest-Paths

In [11] we show how to compute single-source shortest paths in undirected graphs with non-negative edge lengths in $O(\sqrt{nm/B}\log n + MST(n,m))$ I/Os, where $MST(n,m)$ is the I/O-cost of computing a minimum spanning tree (e.g, $\mathrm{sort}(n+m)$ I/Os using a randomized approach). For sparse graphs, the new algorithm performs $O((n/\sqrt{B})\log n)$ I/Os. This result removes our previous algorithm's [10] dependence on the edge lengths in the graph. The new bound is obtained by a number of new ideas to implement a recursive shortest-path algorithm that uses a specific partition into "well-separated" subgraphs, allowing the computation of shortest paths in the whole graph using nearly independent computations on these subgraphs.

Heuristics for Directed Graphs

We continued to work on semi-external DFS: a first parallelization [4] of our DFS approach [13] is currently extended to be used within STXXL [6]. Semi-external DFS is also a key subroutine in order to conduct research on web graphs [7]: We studied a large crawl of 200 million pages and about 1.4 billion edges, and synthetic graphs obtained by the large scale simulation of stochastic graph models for the Webgraph. In particular, we examined the distributions of vertex in-degrees and out-degrees, the PageRank distribution of the vertices, the correlation between the in-degree and the PageRank distribution, and the number of disjoint bipartite cliques in the graph.

Another problem we considered is that to compute the cycle structure of large directed graphs where each node has exactly one outgoing edge. Such graphs appear as state diagrams of finite state machines such as pseudo-random number generators in cryptography. The size of the graphs necessitates that the adjacency list is kept on hard disks. Our heuristic in [5] uses multiple processing units, so that a parallel storage system has to be employed to store the graph. We present experimental results for randomly chosen graphs, and for the graph of the A5/1 generator used in GSM mobile phones. Recently we also developed an alternative approach with improved worst-case I/O-complexity, which is currently being implemented.

I/O-Efficient Well-Separated Pair Decomposition and its Applications

In [8] we present an external memory algorithm to compute a well-separated pair decomposition (WSPD) of a given point set P in $\mathbb{R}^d$ in $O(\mathrm{sort}(N))$ I/Os using $O(N/B)$ blocks of external memory, where N is the number of points in P, and $sort(N)$ denotes the I/O complexity of sorting N items; the dimension d is assumed to be fixed. We also show how to dynamically maintain the WSPD in $O(\log_B N)$ I/Os per insert or delete operation using $O(N/B)$ blocks of external memory. As applications of the WSPD, we show how to compute a linear size t-spanner for P within the same I/O and space bounds and how to solve the K-nearest neighbor and K-closest pair problems in $O(\mathrm{sort}(KN))$ and $O(\mathrm{sort}(N+K))$ I/Os using $O(KN/B)$ and $O((N+K)/B)$ blocks of external memory, respectively. Using the dynamic WSPD, we show how to dynamically maintain the closest pair of P in $O(\log_B N)$ I/Os per insert or delete operation using $O(N/B)$ blocks of external memory.

References

[1] A. Aggarwal and J. S. Vitter. The input/output complexity of sorting and related problems. *Communications of the ACM, 31(9)*, pp. 1116–1127, 1988.

• [2] D. Ajwani, R. Dementiev, and U. Meyer. A computational study of external memory BFS algorithms. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'06*, Miami, USA, 2006, pp. 601–610. ACM-SIAM.

• [3] D. Ajwani, U. Meyer, and V. Osipov. Improved external memory BFS implementations. In D. Applegate and G. Brodal, eds., *9th Workshop on Algorithm Engineering and Experiments (ALENEX)*, New Orleans, USA, 2007. SIAM.

[4] A. Beckmann. Parallelizing semi-external depth first search. Master's thesis, Martin-Luther-Universität Halle-Wittenberg, October 2005.

[5] A. Beckmann and J. Keller. Parallel-external computation of the cycle structure of invertible cryptographic functions. In *15th EUROMICRO International Conference on Parallel, Distributed and Network-Based Processing (PDP'07)*, Naples, Italy, 2007, pp. 526–533. IEEE.

• [6] R. Dementiev, L. Kettner, and P. Sanders. STXXL: Standard template library for XXL data sets. In G. S. Brodal and S. Leonardi, eds., *Algorithms - ESA 2005, 13th Annual European Symposium (ESA 2005)*, Palma de Mallorca, Spain, October 2005, *LNCS 3669*, pp. 640–651. Springer.

• [7] D. Donato, L. Laura, S. Leonardi, U. Meyer, S. Millozzi, and J. F. Sibeyn. Algorithms and experiments for the webgraph. *Journal of Graph Algorithms and Applications*, 10(2), 2007.

[8] S. Govindarajan, T. Lukovszki, A. Maheshwari, and N. Zeh. I/O-efficient well-separated pair decomposition and its applications. *Algorithmica*, 45(4):585–614, August 2006.

[9] K. Mehlhorn and U. Meyer. External-memory breadth-first search with sublinear I/O. In R. Möhring and R. Raman, eds., *Algorithms - ESA 2002, 10th Annual European Symposium*, Rome, Italy, September 2002, *LNCS 2461*, pp. 723–735. Springer.

[10] U. Meyer and N. Zeh. I/O-efficient undirected shortest paths. In G. Di Battista and U. Zwick, eds., *Algorithms - ESA 2003: 11th Annual European Symposium*, Budapest, 2003, *LNCS 2832*, pp. 434–445. Springer.

• [11] U. Meyer and N. Zeh. I/O-efficient undirected shortest paths with unbounded edge lengths. In Y. Azar and T. Erlebach, eds., *Algorithms - ESA 2006, 14th Annual European Symposium*, Zurich, Switzerland, September 2006, *LNCS 4168*, pp. 540–551. Springer.

[12] K. Munagala and A. Ranade. I/O-complexity of graph algorithms. In *Proc. 10th Ann. Symposium on Discrete Algorithms (SODA)*, 1999, pp. 687–694. ACM-SIAM.

[13] J. F. Sibeyn, J. Abello, and U. Meyer. Heuristics for semi-external depth first search on directed graphs. In *SPAA 2002, Fourteenth Annual ACM Symposium on Parallel Algorithms and Architectures*, Winnipeg, Canada, August 2002, pp. 282–292. ACM.

12.6.2 Pass Efficient Algorithms

Investigator: Kevin Chang in collaboration with Ravi Kannan

An important model for massive data set computation in the theoretical computer science literature is the streaming model of computation, in which the input is assumed to be in a read-only array that may be accessed only through a small number of sequential passes. The

algorithm is allotted a small amount of extra space to process the input during each pass. If the read-only array corresponds to data in external storage and extra space corresponds to main memory, this model satisfies the most important constraints imposed by massive data set computation: the data must be in external storage since it is way too large to fit in memory, and frequent random access to external storage is prohibitively expensive.

In the multiple pass-streaming model, the most important resources are the number of passes and amount of memory required by the algorithm. We have studied the tradeoff between passes and memory: Can allocating more passes to the algorithm significantly improve its memory complexity? Can this be proved to be tight?

In the streaming model, we have studied algorithms for data mining and machine learning. Other streaming algorithms of a statistical nature in the computer science literature include the algorithm of Guha *et al* [3].

1. Previously, in collaboration with Ravi Kannan [2], we considered a statistical learning problem, related to generative models of clustering. In this problem, the input is a datastream of samples, placed in arbitrary order, and drawn according to a probability distribution called a *mixture of uniform distributions in* $\mathbb{R}$. The algorithm must learn the density function of the distribution from the samples. Our algorithm has a strong tradeoff between passes and memory: a few more passes allows the algorithm to use much less memory.

2. In [1], we generalized the above algorithm to the high dimensional problem of learning a mixture of uniform distributions over cells in $\mathbb{R}^d$.

3. We are currently studying mathematical programming problems related to statistical computations that are natural for massive data sets, in particular optimization algorithms for parameter estimation of models by maximum likelihood.

References

- [1] K. Chang. Multiple pass streaming algorithms for learning mixtures of distributions in r^d. Full version. Extended abstract submitted to conference., January 2007.
- [2] K. L. Chang and R. Kannan. The space complexity of pass-efficient algorithms for clustering. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'06*, Miami, USA, January 2006, pp. 1157–1166. ACM-SIAM. This is the conference version. Full version under submission to journal.

[3] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *SODA*, 2006, pp. 733–742.

12.6.3 Lock-Free Algorithms for System Services

Investigators: Anders Gidenstam in collaboration with Marina Papatriantafilou, Hakan Sundell, and Philippas Tsigas

During the recent years true processor level parallelism has reached servers, many desktop machines and even laptops thanks to the processor manufacturers' focus on multicore processors. In such multicore and multiprocessor systems the processors have shared access to the

system's main memory and use operations on the memory to communicate and synchronize. In system services, such as scheduling, resource allocation and inter-thread communication, and in concurrent applications on concurrent shared-memory systems efficient synchronization methods are important. In particular, in multicore and multiprocessor systems certain synchronization methods, such as lock-based ones, may limit the parallelism. It is therefore significant to see the impact of lock-free and wait-free synchronization design, which avoids the weaknesses of lock-based designs, in key services for multiprocessor systems.

We have developed lock-free implementations of two such key services: a lock-free memory allocator NBmalloc [3], and a thread library LFthreads [1], whose synchronization is entirely based on lock-free techniques. For both services experiments with benchmark applications indicate that the scalability properties are enhanced further with the help of the lock-free synchronization.

The availability of lock-free data-structures and lock-free utility libraries are important both to make it easier to apply lock-free techniques in applications as well as building blocks in the implementation of lock-free system services. In [2] we presented an efficient and practical lock-free implementation of a memory reclamation scheme based on reference counting, aimed for use with arbitrary lock-free dynamic data structures. Experimental results indicate significant performance improvements for lock-free data structures that require strong garbage collection support.

We are currently working on the full versions of the above mentioned conference papers.

We are also interested in assembling a library of lock-free data-structures as well as support algorithms, such as memory reclamation schemes, for implementing lock-free data-structures and to investigate how to design clean and easy to use and understand interfaces to these components.

References

[1] A. Gidenstam. *Algorithms for synchronization and consistency in concurrent system services.* PhD thesis, Chalmers University of Technology, Göteborg, Sweden, September 2006.

[2] A. Gidenstam, M. Papatriantafilou, H. Sundell, and P. Tsigas. Efficient and reliable lock-free memory reclamation based on reference counting. In *Proceedings of the 8th International Symposium on Parallel Architectures, Algorithms and Networks*, Las Vegas, USA, 2005, pp. 202–207. IEEE.

[3] A. Gidenstam, M. Papatriantafilou, and P. Tsigas. Allocating memory in a lock-free manner. In G. S. Brodal and S. Leonardi, eds., *Proceedings of the 13th Annual European Symposium on Algorithms (ESA'05)*, Palma de Mallorca, Spain, October 2005, *LNCS 3669*, pp. 329–342. Springer.

12.7 Information Retrieval

Coordinator: Holger 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 of what is nowadays called *algorithm engineering*: identify core problems in important real-world applications, analyze these core problems with as much mathematical rigor as possible, complement and/or support these findings by extensive experiments, and build actual systems that demonstrate that indeed the core problems have been identified and the solutions are of practical relevance.

We are a subgroup of the Algorithms and Complexity group, with two PhD students who are about to finish their theses, and five master students; four other master students have finished their theses in the last two years. We interact closely both with other researchers in the Algorithms and Complexity department (D1) as well as with the Information and Databases department (D5). In particular, Gerhard Weikum and Holger Bast are the local coordinators of a large EU project (DELIS). Several joint papers have been produced together with D5 in the last two years, among which is a VLDB paper and a SIGIR paper (still under review). We have also started cooperating with the new Machine Learning Group (R2), headed by Tobias Scheffer.

12.7.1 The CompleteSearch Engine

CompleteSearch is the name of our prototype system, which serves as a testbed and demonstrator for many of our insights and results. CompleteSearch is a fully functional search engine with a web interface, which can be accessed via any standard web browser. More specifically, the system serves the following purposes:

The proof of the pudding: The ultimate proof of the soundness and usefulness of a research result is to integrate it into a real system and check whether the abstract solutions indeed lead to the desired improved behavior in practice. Every single work of ours has been influenced by the feedback we got from this integration process. For example, the formulation of the *context-sensitive range searching problem* underlying the index data structures discussed in Section 12.7.2 were inspired by playing around with an early prototype of our system.

User feedback: We have built a number of concrete applications of CompleteSearch and made them publically available [1]. The feedback from our users, which includes ourselves, is an important source of inspiration for new research problems. Moreover, these applications provide us with real query logs.

Demonstrator: There is nothing more convincing than a (working) live demonstration. CompleteSearch has already been presented at a variety of conferences, universities, and companies (including Yahoo and Google). The feedback was consistently extremely positive and encouraging.

Commercial utilization: A number of business contacts have arisen due to live demonstrations we have given or due to the publicly available demos on our website. Several commercial applications are currently being negotiated. CompleteSearch is soon to become the official search engine of the DBLP server; a prototype is already available at `http://search.mpi-inf.mpg.de/dblp`.

The various features of the current system, and the research work behind them, will be explained in the following subsections. From a high-level perspective, the system consists of four parts: (1) an index builder and associated tools, written in C++ and perl; (2) a query processor, written in C++; (2) the web server application, written in PHP; (3) the user

interface, written in HTML with embedded JavaScript. Each of these parts have become large and complex pieces of software, and the interaction between them can be very complex and hard to debug (asynchronous, networking, multithreaded, etc.).

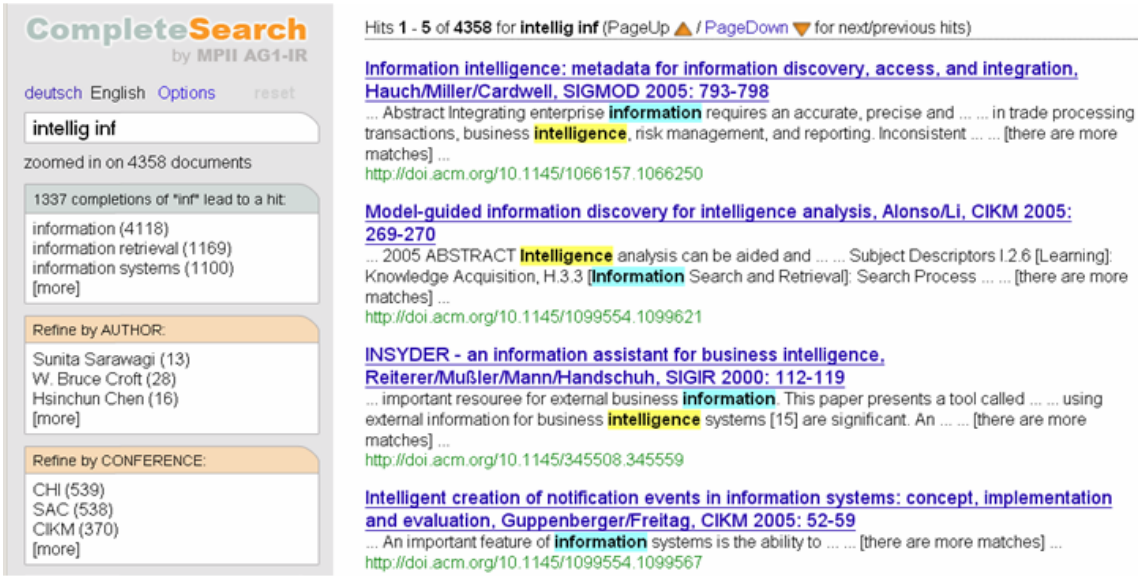


Figure 12.9: A screenshot of the CompleteSearch engine in action for the query `intellig inf` on a collection of computer science articles, indexed with full text + meta data. The list to the right shows documents which contain both a word starting with `intellig` and a word starting with `inf`. The first box below the search field shows words, subwords, and phrases starting with `inf` that lead to the best hits. The other boxes show a breakdown of the whole set of 2302 hits by author and conference. A key fact to note is that the various features are all provided via one and the same, carefully designed and implemented prefix search and completion mechanism.

References

- [1] The CompleteSearch engine: Applications and live demonstrations. http://search.mpi-inf.mpg.de.

12.7.2 Context-sensitive Range Search and Index Building

Investigators: Holger Bast, Debapriyo Majumdar, Christian Worm Mortensen, and Ingmar Weber

The standard data structure behind most, if not all big search engines is the *inverted index*. In its simplest form, this entails precomputing, for each word that occurs somewhere in the given collection of documents, the sorted list of (ids of) documents containing that word.

The inverted index combines many properties which are invaluable for large-scale retrieval: the index lists can be compressed very well (and so more data can be read from disk in less

time), access is very IO-efficient (lists are stored contiguously on disk), it can be constructed relatively fast (essentially a sort), and it is easy to implement and extend [3].

Its main drawback is its limited use: an inverted index is perfect for keyword search as we know it from the web search engines like Google, but it cannot be used advantageously for more complex operations as we know them from database management systems, such as selects and joins on tables.

In [2] and [1] we have devised more powerful index data structures which have all the good properties from an inverted index, listed above, but can solve the following general-purpose *range searching problem*:

Given a set D of documents, and a range W of words, compute all pairs (w, d) such that $w \in W$, $d \in D$, and w occurs in document d.

Many powerful search features can be reduced to this range searching problem: context-sensitive prefix search aka autocompletion (see Figure 12.10), faceted search (see Section 12.7.3), XML and semantic search (see Section 12.7.5), database-style queries (see Section 12.7.4, and more. Standard keyword search is contained as a special case ($|W| = 1$).

The algorithm from [1] is output-sensitive in that its running time is linear in the size of the input plus the size of the output. It is space-efficient in that it uses no more space than an *uncompressed inverted index*. However, due to its heavy use of bit-rank data structures (for counting the number of 1s up to a given position in a large bit vector), the algorithm is inherently not particularly IO-efficient.

The algorithm from [2] is tuned towards dealing with very large amounts of data. It makes heavy use of compression and provably uses no more space than a *compressed inverted index*. And it is prefect in terms of IO-efficiency. However, it is no longer output-sensitive.

We are currently working on an index data structure and query algorithm that is highly compressible and IO-efficient, and also output-sensitive.

References

- [1] H. Bast, C. W. Mortensen, and I. Weber. Output-sensitive autocompletion search. In F. Crestani, P. Ferragina, and M. Sanderson, eds., *String Processing and Information Retrieval, 13th International Conference, SPIRE 2006*, Glasgow, GB, 2006, *LNCS 4209*, pp. 150–162. Springer.
- [2] H. Bast and I. Weber. Type less, find more: Fast autocompletion search with a succinct index. In E. N. Efthimiadis, S. Dumais, D. Hawking, and K. Järvellin, eds., *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, USA, 2006, pp. 364–371. ACM.

[3] I. H. Witten, T. C. Bell, and A. Moffat. *Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd edition.* Morgan Kaufmann, 1999.

12.7.3 Faceted Search

Investigators: Holger Bast and Ingmar Weber

Faceted search refers to the following useful and convenient search feature, which by now has become standard in e-shopping applications like Ebay and Amazon. Assume each document is associated with a number of labels, and labels belong to different categories. For example, assume that in a collection of computer science papers, each document is associated with the

names of its authors (all belonging to the category "author"), with the year of publication (belonging to the category "year"), and with the name of the conference (belonging to the category "conference").

In faceted search, we then have, along with the search results for a query, also the most significant (frequent) labels occurring in that result set, grouped by category (facet). The "refine by author" and "refine by conference" boxes in Figure 12.9 give an example.

More than that, at every point in the search the user now has *two* options to further refine the search: (i) to add another query word, as in ordinary keyword search; this will restrict the current set of documents to those which also contain the newly added keyword; and (ii) to select a particular label; this will restrict the current set of documents to those which have this label attached.

Research on faceted search has so far concentrated on the (important) user-interface aspect [3] [2]. There has been hardly any work on the efficiency aspect.

In [1] we have contributed to both these aspects. We have suggested a new interactive variant of faceted search, and we have shown that it can be implemented efficiently using one of our data structures described in Section 12.7.2.

References

- [1] H. Bast and I. Weber. When you're lost for words: Faceted search with autocompletion. In A. Broder and Y. Maarek, eds., *SIGIR'06 Workshop on Faceted Search*, Seattle, USA, August 2006, pp. 31–35. ACM.

[2] J. English, M. Hearst, R. Sinha, K. Swearingen, and K.-P. Yee. Hierarchical faceted metadata in site search interfaces. In *Conference on Human factors in computing systems (CHI'02)*, 2002, pp. 628–639.

[3] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Yee. Finding the flow in web site search. *Communications of the ACM*, 45(9):42–49, 2002.

12.7.4 DB & IR Integration

Investigators: Holger Bast and Ingmar Weber

Search engines scale very well to huge amounts of data, but they are very special purpose: they are tuned towards the specific application of keyword search and that is essentially all they can be used for.

Database management systems, on the other hand, are very general-purpose: they can handle complex selects and joins on arbitrary tables, they can handle dynamic updates, there is concurrency control and recovery, etc. The price is that these systems do not scale well to huge amounts of data, they either use a lot of space or tend to be slow, or both.

Bridging this gap is a hot research topic [2], and of great practical relevance for companies like Yahoo and Google who have massive amounts of both text data (web pages, emails, etc.) and structured data (user accounts, shopping articles, etc.). Several novel databases systems making extensive use of compression and locality of access (the two major factors which make search engines so scalable) are currently being developed and marketed [3].

In [1], we show how our index data structures from Section 12.7.2 can be used to provide a number of standard database operations efficiently, in particular selects and joins on arbitrary

tables. In this work, we also give an overview of our CompleteSearch engine from the system's perspective and give an account of some of the lessons that we have learned in our two years of work on that project.

References

- [1] H. Bast and I. Weber. The completesearch engine: Interactive, efficient, and towards ir & db integration. In G. Weikum, ed., *3rd Biennial Conference on Innovative Data Systems Research (CIDR'07)*, Asilomar, USA, 2007. VLDB Endowment.

[2] S. Chaudhuri, R. Ramakrishnan, and G. Weikum. Integrating DB and IR technologies: What is the sound of one hand clapping? In *2nd Biennial Conference on Innovative Data Systems Research (CIDR'05)*, 2005, pp. 1–12.

[3] M. Zukowski, P. A. Boncz, N. Nes, and S. Héman. MonetDB/X100 - a DBMS in the CPU cache. *IEEE Data Engneering Bulletin*, 28(2):17–22, 2005.

12.7.5 Semantic Search

Investigators: Holger Bast, Alexandru Chitea, Fabian Suchanek, and Ingmar Weber

The current generation of search engines does ranked keyword search: for a given query, they return a list of documents, ordered by relevance, which contain some or all of the query words. Ranked keyword search has been quite successful in the past, for popular queries in web search, but also in scientific benchmarks like those of TREC [2].

However, keyword search has its obvious limits, and there is no doubt that the next generation of search engines will be more "semantic" in one way or the other. For example, consider the query "which musicians are associated with The Beatles". This requires a search not for the literal *word* musician, but rather for *instances of the class* it denotes.

Already this simple query highlights two of the main challenges of semantic search: (1) obtain the necessary semantic information, in this case, identify each occurrence of a musician in the given text collection; and (2) make that information searchable in a convenient and efficient way.

In [1], we present *Ester*, a modular and highly efficient system for combined full-text and ontology search. We show how Ester can answer arbitrary SPARQL 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 context-sensitive 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. As a proof of concept, we applied Ester to the English Wikipedia, which contains about 2 million documents, combined with the Yago ontology developed at D5 [3], 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 on text, entities, and relations, 2007. submitted to SIGIR'07.

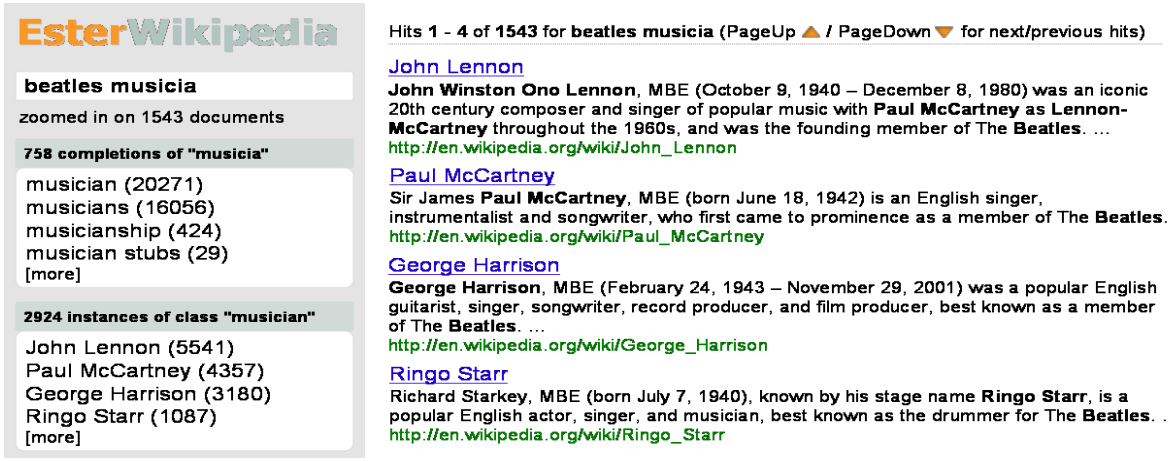


Figure 12.10: A screenshot of our search engine for the query `beatles musicia` searching the English Wikipedia. The list of completions and hits is updated automatically and instantly after each keystroke, hence the absence of any kind of search button. The number in parentheses after each completion is the number of hits that would be obtained for that particular completion. The upper box suggest words and phrases that start with `musicia` and that occur together with the word `beatles`. The lower box suggests *instances of musicians* that occur together with the word `beatles`. Fast processing of this apparently simple query requires the whole complexity of our system in the background: prefix queries, join queries, entity recognition, and ontological knowledge.

[2] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *4th Text Retrieval Conference (TREC'95)*, 1995, pp. 73–96.

- [3] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *16th international World Wide Web conference (WWW 2007)*, 2007.

12.7.6 User Study

Investigators: Holger Bast and Ingmar Weber

The major motivation for the development of CompleteSearch was and is to find information fast and effectively. In [1] we tested this hypothesis via a user study with a concrete application, namely the helpdesk of our institute. In that work, we also discussed a relation of our system to *case-based reasoning* [2].

The MPII helpdesk offers first-level computer support to all users in the institute. Helpdesk employees search on a daily basis in a database of over 7,000 former help requests to find if a similar problem existed in the past and, if so, if and how it was solved. For our user study, we provided searching and browsing facility via two concrete systems: (A) Google Desktop Search, a standard off-the-shelf search software, and (B) our CompleteSearch engine.

Five users where given a set of 10 problems each. For each of these problems they were given 2 minutes to find relevant information in the document collection. All of the five users used system A for half the problems and system B for the other half. Furthermore, we

ensured that for each problem one half of the test users used system A and the other half used system B.

The results were overwhelmingly positive in favor of CompleteSearch: (i) the test users found it easier to find relevant information with CompleteSearch, and (ii) the users also preferred the search experience of our system. All test users unanimously mentioned the system's speed as its strongest point. Our helpdesk is now using our CompleteSearch system on a daily basis.

References

- [1] H. Bast and I. Weber. Managing helpdesk tasks with completesearch: A case study. In B. Decker and M. Nick, eds., *4th Conference on Professional Knowledge Management (WM'07)*, Berlin, Germany, 2007, LNCS. Springer.

[2] R. Bergmann, K.-D. Althoff, S. Breen, M. Göker, M. Manago, R. Traphöner, and S. Wess. *Developing Industrial Case-Based Reasoning Applications.* Springer-Verlag GmbH, second edition, 2004.

12.7.7 Top-k Query Processing

Investigators: Holger Bast, Debapriyo Majumdar, Martin Theobald, Ralf Schenkel, and Gerhard Weikum

Top-k query processing is an important building block for ranked retrieval, as the users look for only the top 10, 20 or at most 50 hits from a search engine.

Top-k queries operate on inverted index, and aim to return the results without accessing the lists corresponding to the terms fully. Among the previous works, the most well known methods are Fagin's NRA and CA algorithms, which perform sequential or sorted access on index lists sorted by score of documents and also use random access to resolve score uncertainty. In spite of being able to return the top-k results without scanning the lists fully (i.e. saving some access cost), these algorithms are not very effective in practice, as a simple algorithm which merges the lists fully and partially sorts the merged list to determine the top-k always beats the more sophisticated top-k algorithms in running time.

In [2], we take a new, principled approach towards scheduling sequential accesses based on a Knapsack related optimization problem and schedule random access by accurately estimating random access cost at runtime. More importantly, we also present a novel data structure called *inverted block index* which enables us to efficiently implement our algorithm. Figure 12.11 shows one instance of our algorithm on a small example.

In performance experiments with three different datasets (TREC Terabyte, HTTP server logs, and IMDB), our methods achieved significant performance gains compared to the best previously known methods: a factor of up to 3 in terms of execution costs, and a factor of 5 in terms of absolute run-times of our implementation. Our best techniques are close to a lower bound for the execution cost of the considered class of threshold algorithms.

We also used our implementation [1] to participate in the Terabyte track of TREC 2006 with very good results (as seen in the overview paper [4]) in the efficient task. In spite of being disk based and run on essentially single and dual processor configurations, all four of our runs were among the first 12 of the 25 submitted runs at TREC. Our fastest run had

List 1	**List 2**	**List 3**
Doc17 : 0.8	Doc25 : 0.7	Doc83 : 0.9
Doc78 : 0.2	Doc38 : 0.5	Doc17 : 0.7
·	Doc14 : 0.5	Doc61 : 0.3
·	Doc83 : 0.5	·
·	·	·
·	Doc17 : 0.2	·
·	·	·

Round 1 (SA on 1,2,3)

Doc17 :	[0.8 , 2.4]
Doc25 :	[0.7 , 2.4]
Doc83 :	[0.9 , 2.4]
unseen:	≤ 2.4

Round 2 (SA on 1,2,3)

Doc17 :	[1.5 , 2.0]
Doc25 :	[0.7 , 1.6]
Doc83 :	[0.9 , 1.6]
unseen:	≤ 1.4

Round 3 (SA on 2,2,3**!**)

Doc17 :	[1.5 , 2.0]
Doc83 :	[1.4 , 1.6]
unseen:	≤ 1.0

Round 4 (RA for Doc17)

Doc17 : 1.7
all others < 1.7
done!

Figure 12.11: A top-1 computation on three index lists, with three rounds of sorted access, followed by one round of random access.

an average query processing time of about 28 milliseconds per query for 100,000 queries of 4 words on average, in spite of being disk based. The only other method [3] which was faster than ours was primarily memory based for the efficiency runs.

References

- [1] H. Bast, D. Majumdar, R. Schenkel, M. Theobald, and G. Weikum. IO-top-k at TREC. In *In 15th Text Retrieval Conference (TREC'06)*, 2006.

- [2] 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.

[3] S. Büttcher, C. L. A. Clarke, and I. Soboroff. The trec 2006 terabyte track (overview paper). In *The Text Retrieval Conference (TREC 2006)*, 2006.

[4] S. Büttcher, C. L. A. Clarke, and P. C. K. Yeung. Index pruning and result reranking: Effects on ad-hoc retrieval and named page finding (wumpus at trec 2006). In *The Text Retrieval Conference (TREC 2006)*, 2006.

12.7.8 Latent Semantic Indexing and Related Methods

Investigators: Holger Bast and Debapriyo Majumdar in collaboration with George Dupret and Benjamin Piwowarski

In [2] we introduced *relatedness curves* as a means to understand and analyze latent semantic indexing (LSI)[3] and related *spectral retrieval* methods. We showed that all existing retrieval methods based on an eigenvector decomposition can be viewed as (implicitly) assigning a relatedness score to each term pair. For a fixed term pair, this score depends on the number of eigenvectors picked, and the relatedness curve is simply the curve of all scores for a growing sequence of eigenvectors.

Our findings in [2] gave strong theoretical and empirical evidence that actual term relatedness is more related with the *shape* of the relatedness curve than with individual values on the curve; see Figure 12.12 for a number of examples. Methods for the fully automatic derivation of term taxonomies had been considered before, however, not in the context of eigenvector decomposition [4].

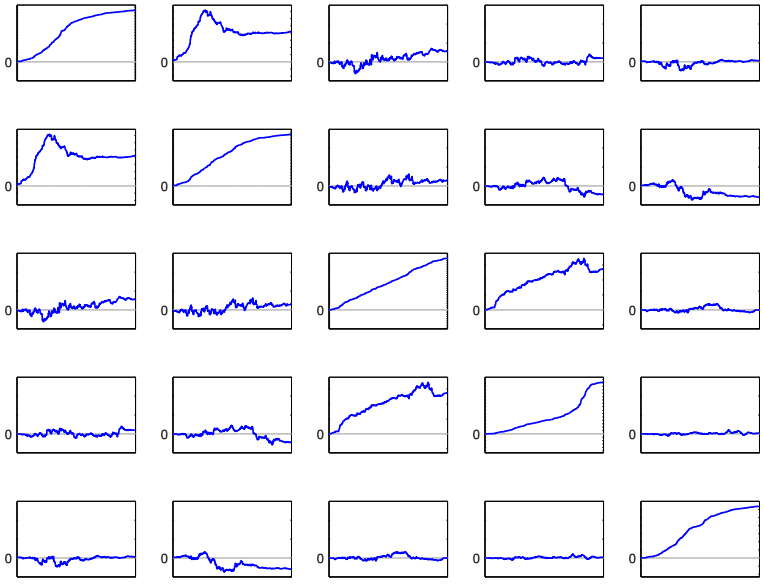


Figure 12.12: All 25 pairwise relatedness curves for all pairs of 5 terms. The first and the second term are related, so are the third and the fourth term, as indicated by both the smoothness and the mononicity of the respective curves. In contrast, the relatedness curve for unrelated term pairs tends to be a random oscillation about zero.

In [2], term-term relations were symmetric. In [1] we extended our mathematical model to show that spectral methods actually have the power to identify asymmetric relation, for example, that “orange” is more specific than the related ”fruit” and not vice versa; more such examples found by our method are given in Table 12.1.

more specific	→	more general	more specific	→	more general
alberta	→	canada	glycerin	→	soap
iguana	→	lizard	salsa	→	sauce
transpersonal	→	psychology	catalyst	→	chemistry

Table 12.1: Asymmetric term-term relations found fully automatically by our method on a collection of 500,000 news articles

References

- [1] H. Bast, G. Dupret, D. Majumdar, and B. Piwowarski. Discovering a term taxonomy from term similarities using principal component analysis. In M. Ackermann, B. Berendt, M. Grobelnik, A. Hotho, D. Mladenic, G. Semeraro, M. Spiliopoulou, G. Stumme, V. Svatek, and M. W. van Someren, eds., *Semantics, web and mining, Joint International Workshops, EWMF 2005 and KDO 2005*, Porto, Portugal, 2006, *LNCS 4289*, pp. 103–120. Springer.

- [2] H. Bast and D. Majumdar. Why spectral retrieval works. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, eds., *28th Annual International Conference on Research and Development in Information Retrieval (SIGIR'05)*, Salvador, Brazil, August 2005, pp. 11–18. ACM.

[3] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[4] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999, pp. 206–213.

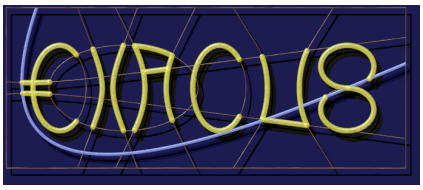
12.8 Geometric Computing

Coordinator: Michael Sagraloff

Software systems and algorithm engineering are categories that cut across the other areas of our group. Some of the efforts are thus documented in the respective area. Similarly, the related theoretical work is covered in the respective sections, such as in Section 12.3 for Foundations and Discrete Mathematics.

We continued ourwork on Exacus, a project we created in April 2002 to study arrangements of curves and surfaces (Section 12.8.1). We made our second public release of Exacus available in August 2006 as part of the EU funded project called Acs, *Algorithms for Complex Shapes*. Within the first year of the reporting period the MPI was leading the corresponding work package with Lutz Kettner appointed as software coordinator in Acs. About half of the resources in the project are dedicated to software development with the main focus on extending Cgal towards curves and surfaces. We continued our work on Cgal (Section 12.8.2) with emphasis on this goal and are integrating important parts of Exacus into Cgal. The goal for the next reporting period is to finalize this integration process.

12.8.1 EXACUS: Efficient and Exact Algorithms for Curves and Surfaces



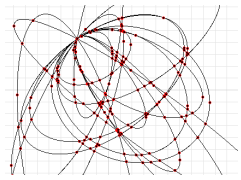
The MPI is participating in the ACS-project (*Algorithms for Complex Shapes with certified numerics and topology*[2]) and is the successor of the ECG-project (*Effective Computational Geometry for Curves and Surfaces*[3]). The ACS-project started in May 2005 and will last until May 2008. It is a cooperation with six European research groups in Groningen, ETH Zürich, FU-Berlin, INRIA Sophia-Antipolis, Athens, Tel-Aviv and GeometryFactory[4], the sole distributor of commercial licenses for CGAL. ACS aims to advance the handling of complex shapes in geometric algorithms, both in theory and practice. Our results contributed to this project.

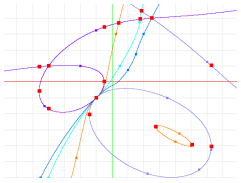
With respect to linear geometry, CGAL is the state-of-the-art in implementing geometric algorithms completely, exactly, and efficiently. When CGAL was started, the research area of Computational Geometry already had existed for two decades. With this long history it was clear how to implement the algorithms for linear geometry robustly and efficiently. However, while many geometric algorithms apply also to non-linear objects the number of degenerate cases grows dramatically when going from straight-line to curved objects. Thus the final goal of a stable software library, with re-usable interfaces, for curved objects was unlikely to be achieved in the first attempt.

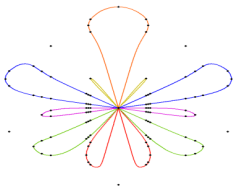
For the reasons discussed above we decided to outsource our investigations for curved geometry from CGAL and conceived the EXACUS-project (*Efficient and Exact and Algorithms for Curves and Surfaces*[5]). It was created in April 2002 in the context of the ECG-project. We continued our work on EXACUS as part of the ACS-project. Although we call our library design prototypical, we spent nonetheless a great effort on completeness, exactness, efficiency, and documentation. The software libraries in EXACUS are designed to support further research implementations based on these libraries. Within the last two years, EXACUS proved to be a good laboratory to improve our ideas for implementations for curves and surfaces. From there and in order to profit from synergy effects, the ACS partners decided to integrate core parts of EXACUS into CGAL. We started the integration process right after the second public release of EXACUS in August 2006. For the upcoming CGAL release 3.3, we in particular accomplish the reorganization of the number type support into the new CGAL package Algebraic Foundations. This will allow us to integrate further parts of EXACUS into CGAL, while CGAL will in turn serve as the main basis of further projects within EXACUS.

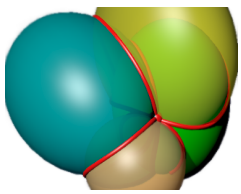
In the context of ACS and EXACUS in the last two years, we have made several contributions to the exact and efficient computation of arrangements of curved objects. An overview of the EXACUS-design is given in the next subsection, followed by more detailed subsections about our investigations on exact computation of arrangements of rotated conics [1], lower envelops of spatial quadrics [4], and analysis of planar curves of general degree [6, 11]. A result on an algorithm for isolating real roots for polynomials with bit-stream coefficients [7] is discussed in Subsection 12.3.3 on polynomial solving of the Foundations and Discrete

[2] http://acs.cs.rug.nl/
[3] http://www-sop.inria.fr/prisme/ECG/
[4] http://www.geometryfactory.com/
[5] http://www.mpi-sb.mpg.de/EXACUS/


ConiX(CnX)


CubiX(CbX)


AlciX(AcX)


QuadriX(QdX)

SweepX (SoX)							
NumeriX (NiX)							
Library Support (LiS)							
Boost	Gmp	NTL	Core	Leda	Cgal	BMTools	Qt

Figure 12.13: Library layers of the Exacus project.

Mathematics Section 12.3.

Software in Exacus

Investigators: Eric Berberich, Arno Eigenwillig, Michael Hemmer, Michael Kerber, Lutz Kettner, Kurt Mehlhorn, Joachim Reichel, Tobias Reithmann, Elmar Schömer, and Nicola Wolpert

Exacus is a collection of C++ libraries, see Figure 12.13 for their layered architecture. At the bottom we see the external libraries that we use in Exacus. Library Support provides the foundation, such as configuration and memory management. NumeriX provides number type support, symbolic algebra, and numerical algorithms. SweepX contains our generic sweep-line algorithm suitable for segments of all type of curve arcs and boolean operations based on it. The top layer illustrates the four applications pursued so far; ConiX, CubiX, and AlciX compute arrangements of curves in the plane, and QuadriX extends our work to spatial arrangements of quadrics.

The libraries consist currently of about 90000 lines of code including the documentation that is embedded in the C++ source code. Our design follows the generic programming paradigm with C++ templates similar to design principles in Cgal. We use Doxygen to create the reference documentation. We use Cvs for version control and distributed collaboration. Configuration and build management is done with the Gnu family of tools `autoconf`, `automake`, and `libtool`. We run daily a fully automatized test-suite from Cvs based on our own scripts.

We detect a couple of other libraries during configuration. Exacus does not depend on these libraries in various core parts of its functionality. However, in other parts, we did not reinvent the wheel and depend on existing implementations. Generic programming allows us to stay flexible and postpone the decision between several alternatives to the final user application code, for example, the choice of number types. For more details about the library design we refer to [2, 13].

We made our second public release[6] of EXACUS available in August 2006. The project is released under the open source license QPL, which is the license also chosen for the CGAL basic library. The release includes the basic Library Support, NUMERIX, SWEEPX, and CONIX. EXACUS has been successfully tested with g++ 3.3 and g++ 4.1. It should also work fine with g++ 3.4 and g++ 4.0. The supported platform is Linux.

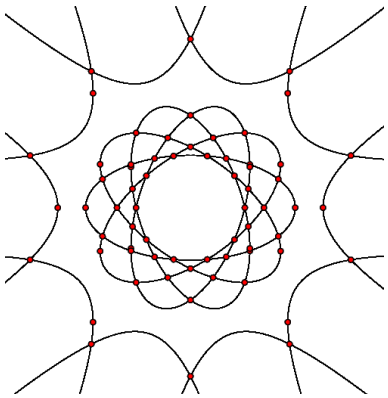
CONIX – Arrangements of Rotated Conics

Investigators: Eric Berberich, Manuel Caroli, and Nicola Wolpert

We extended the CONIX library of EXACUS to compute the arrangement of rotated conics by angle that are constructible with the help of straight-edge and compass [1], i.e., whose sin and cos are representable by nested square-root expressions. Important examples of possible rotation angles are multiples of 45°, 30°, and 15°. A conic is an algebraic curve that is defined by the zero set of a bivariate polynomial of degree 2. The arrangement of transformed conics is the subdivision of the plane induced by them into cells of dimension 0 (*vertices*), of dimension 1 (*edges*), and of dimension 2 (*faces*). Transformations of geometric objects, like translation and rotation, are fundamental operations in CAD-systems. Rotations, in general, trigger the need to deal with trigonometric functions, which is hard to achieve when aiming for exact and robust implementation. Former approaches [5] had to approximate such angles by angles whose sin and cos are rational. Each desired angle can then be arbitrarily approximated by existing implementations for arrangements of conics. But this method leads to increased bitlengths of the involved coefficients and still does not compute the exact solutions as expected by the *exact geometric computation* paradigm [20]. Our generic design allows to reuse as much code as possible. The main problem one has to solve to rotate conics by the mentioned angles is root-isolation of univariate polynomials $p(x) \in \mathbb{K}_{rot}[x]$ where $\mathbb{K}_{rot} = \mathbb{K}_n$ is a normal field extension of $\mathbb{Q}$ with degree 2^n, i.e., there exists a sequence $\mathbb{K}_0 = \mathbb{Q}$, $\mathbb{K}_{i+1} = \mathbb{K}_i(\sqrt{r}), r \in \mathbb{K}_i$. To perform this task we can either use a modified version of the Descartes method or, if $n = 1$, a specialized method, that infers the isolating intervals from roots of integral polynomials.

To compute the arrangement of rotated conics we extended the EXACUS libraries by the two mentioned real root isolators. We also introduced a new representation class template for transformed conics (`CnX::Rotated_conic_2`), that provides access to a *transformation history*, i. e., a sequence of rotations and translations applied to the original integral polynomial. Its actual behavior is determined by a model of the *RotatedConicTraits* concept. It especially defines the allowed rotation angles, the number type used for the coefficients of the bivariate polynomial, and the method to isolate the real roots of its univariate counterparts. For the number type of the coefficients we rely on `NiX::Sqrt_extension`. It represents (nested) square root extensions, i. e., numbers of the form $a + b \cdot \sqrt{c}$, where a, b, and c are of type `Integer` or, if nested, even another instance of `NiX::Sqrt_extension` again. All our mathematical algorithms are designed such that the number type can easily be exchanged from `Integer` to the needed `NiX::Sqrt_extension`. A such equipped version of `CnX::Rotated_conic_2` inherits all the functionality from the generic `CnX::Conic_2`

[6] `http://www.mpi-inf.mpg.de/projects/EXACUS/downloads.html`

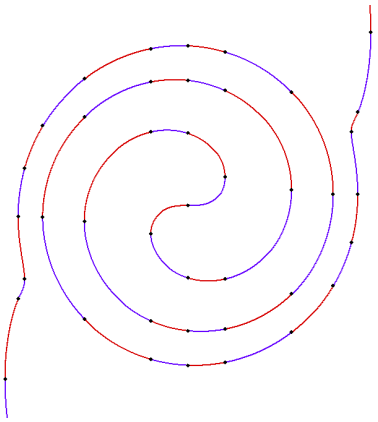


class which is required for arrangement computations (e. g., with CGAL's `Arrangement_2` package) or even to compute boolean set operations on polygons [19] bounded by arcs of rotated conics. The latest release of EXACUS, version 1.0, contains traits classes to rotate conics by multiples of 45°, 30°, and 15°, dealing with all degenerate cases. Rotations by other angles constructible with straight edge and compass can be implemented straightforward, as we recently did for 36°. The figure to the right shows the example of an arrangement of a hyperbola and an ellipse, both rotated by angles of 0°, 36°, 72°, 108°, and 144°.

ALCIX – Algebraic Curves In EXACUS

Investigators: Eric Berberich, Arno Eigenwillig, Pavel Emiliyanenko, Michael Kerber, Michael Sagraloff, and Nicola Wolpert

The goal of the ALCIX library is an efficient implementation of arrangements of algebraic curves defined by polynomials of arbitrary degree. Work on ALCIX has started in February 2006. It currently consists of about 11000 lines of code.

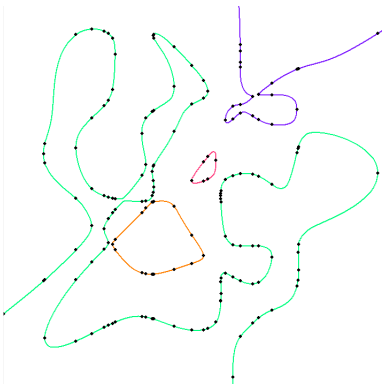


So far, the ALCIX library contains functionality for the analysis of one algebraic plane curve. More precisely, it provides a representation of the curve that allows to query certain topological and geometric properties such as the position of singularities. In particular, the curve can be decomposed into x-monotone segments which are suitable as input for sweep-line algorithms (compare the picture on the right, the segments are drawn in alternating colors).

ALCIX implements a complete and efficient algorithm for analyzing these plane curves [6, 11], extending earlier work of our group on curves of bounded degree [8, 3]. It is complete because it can handle all sorts of degeneracies, such as covertical critical points, vertical asymptotes or any type of singularity. In algorithms dealing with algebraic objects, the use of symbolic computation is necessary to guarantee correct results, but this tends to slow down the analysis; ALCIX is efficient because it reduces the amount of such symbolic computations and replaces them by validated numeric operations. The *Bitstream Descartes method* (see Subsection 12.3.3) is of particular relevance because it allows to solve the important subproblem of real root isolation with purely numerical methods. To handle also polynomials with multiple roots (which naturally arise when dealing with algebraic curves), symbolic precomputation is necessary to compute additional information about the polynomial, such as the number of real roots. For that, we use an algebraic tool called *Sturm-Habicht* sequences which are closely related to *s*ubresultants. Their computation is one of the bottlenecks in the algorithm; we investigated how the two main methods for subresultant computation (pseudo-division-based versus determinant-based) behave in the special case of multivariate polynomials [12].

Using numerical methods to speed up the analysis of curves has been investigated more or less extensively. However, all approaches so far either cannot guarantee exactness of the outcome (e.g. [10]), or they fail in degenerated situations (e.g. [18]). If a failure occurs, a completely symbolic (and slower) method is applied. To our knowledge, our approach is the first that profits from numerical methods in each instance: One moves away from problematic, degenerated situations with a linear change of coordinates, but the coordinates are changed back to the original system eventually.

The algorithm has been compared to the algorithms `top` [10] and `insulate` [17] which compute only the topology of an algebraic curve. AlciX has shown the best performance among those three approaches for the most test examples. We also compared it to `cad2d` from Brown. This is an optimized variant of `qepcad`[7] (a system for quantifier elimination and cylindrical algebraic decomposition) for the case of plane curves. The experiments show that AlciX can handle cases which are infeasible for `cad2d`, especially for singular curves of high degree and polynomials with big coefficients.



Based on the curve analysis, AlciX contains a program to produce reliable plots of the curve. It also allows to draw any subset of x-monotone segments efficiently. This approach given in [9] is a composite of ideas taken from space subdivision and curve tracking methods equipped with techniques facilitating separation of closely located curve branches. It operates on distinct segments delimited by end-points which can also lie at infinity, and employs pixel techniques to always stay within the level of detail given by the current pixel resolution and omit drawing of potentially imperceptible parts. To provide a correct and reliable visualization the algorithm offers a three-level model to gradually increase arithmetic precision in case of need. The picture on the right shows a plot of a curve of degree 10, produced with this program.

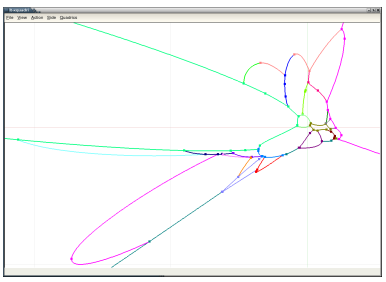
The analysis of a pair of algebraic curves (this includes finding their intersection points as a main step) is implemented in AlciX in a prototypical form so far, an efficient implementation is in the focus of current research. Using the generic sweep-line mechanism of Exacus [2, 8] this curve pair analysis directly leads to an arrangement algorithm for algebraic curves.

QuadriX – Computing Envelopes of Quadrics

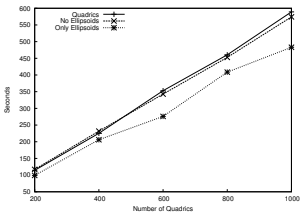
Investigator: Eric Berberich in collaboration with Michal Meyerovitch

In collaboration with the University of Tel Aviv, we implemented a traits class to compute envelopes of quadrics [4]. A quadric is an algebraic surface that is defined by the zero set of a trivariate polynomial of degree 2. Quadrics are at the front of research for non-linear surfaces in three-dimensional modelling [3, 14, 16]. The main difficulty in computing arrangements of quadrics exactly is that high-degree algebraic numbers are involved, even when the quadrics are defined by rational coefficients. An important task is to robustly deal with the three-dimensional intersection curves of quadrics.

[7] `http://www.cs.usna.edu/~qepcad/B/QEPCAD.html`



Lower envelopes are fundamental structures in computational geometry, which have many applications like computing general Voronoi diagrams, or performing hidden surface removal. Its output consists of a minimization diagram which is defined as the point-wise minimum of partially defined functions. The upper envelope uses the maximum of such functions respectively. In the case of quadrics the lower parts of them define these partial function in two variables. Several algorithms exist to compute lower envelopes. We rely on recent work of Meyerovitch who presented a generic and exact implementation of a divide-and-conquer algorithm in CGAL that decouples the combinatorial part from the geometric predicates [15]. In order to support a new class of surface patches for this algorithm it suffices to implement a set of geometric types and operations on them. We had to provide basic surface types: `Surface_3` and `Xy_monotone_surface_3` are both mapped to `QdX::Quadric_3`. In addition, two methods are responsible to compute the projection of surface patches and of the intersections of two surface patches to the plane. These tasks benefit from the existing implementation in the QUADRIX library of EXACUS. The traits class also expects to determine the relative z-order of two xy-monotone surface patches with respect to some projected geometric objects. It distinguishes five methods for such comparisons, all of which use ray-shooting in the z-direction as a basic technique. For the sake of simplicity we only mention the comparison over a projected intersection curve and the comparison over a projected point. While it is possible to use a rational point next to a projected intersection curve to determine the correct relative z-order, in the case of the comparison over a point we have to do the hard work. This means to convert the coordinates of the point to `leda::real` or `CORE::Expr` and rely on their newly introduced ROOTOF methods. Full details are explained in [4]. The given pictures shows the minimization diagram of a mixed set of 400 ellipsoids and hyperboloids.



We also measured the running time of the divide-and-conquer algorithm when computing lower envelopes of quadrics. For increasing n we generated five sets of n random quadrics whose coefficients are ten-bit integers. We further distinguish between mixed quadrics, ellipsoids only, and sets that do not contain ellipsoids. The figure to the right shows the resulting running times on a 3 GHz Pentium IV machine with 2 MB of cache, averaged over all the sets of the same size. Note that computing the lower envelope of 1000 quadrics takes less than 10 minutes, using *exact* arithmetic of LEDA. Computing the lower envelope of ellipsoids is faster.

References

- [1] 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, March 2007, pp. 231–234. Technische Universitaet Graz.
- [2] E. Berberich, A. Eigenwillig, M. Hemmer, S. Hert, L. Kettner, K. Mehlhorn, J. Reichel, S. Schmitt, E. Schömer, and N. Wolpert. Exacus: Efficient and exact algorithms for curves

and surfaces. In G. S. Brodal and S. Leonardi, eds., *13th Annual European Symposium on Algorithms (ESA 2005)*, Palma de Mallorca, Spain, October 2005, *LNCS 3669*, pp. 155–166. Springer.

- [3] E. Berberich, M. Hemmer, L. Kettner, E. Schömer, and N. Wolpert. An exact, complete and efficient implementation for computing planar maps of quadric intersection curves. In J. Mitchell, G. Rote, and L. Kettner, eds., *21st Annual Symposium on Computational Geometry (SCG'05)*, Pisa, Italy, June 2005, pp. 99–106. ACM.

- [4] E. Berberich and M. Meyerovitch. Computing envelopes of quadrics. In *Proceedings of 23rd European Workshop on Computational Geometry*, Graz, Austria, March 2007, pp. 235–238. Technische Universitaet Graz.

[5] J. Canny, B. Donald, and E. K. Ressler. A rational rotation method for robust geometric algorithms. In *SCG '92: Proceedings of the eighth annual symposium on Computational geometry*, New York, NY, USA, 1992, pp. 251–260. ACM Press.

- [6] A. Eigenwillig, M. Kerber, and N. Wolpert. Fast and exact geometric analysis of real algebraic plane curves. Accepted for the International Symposium on Symbolic and Algebraic Computation (ISSAC 2007), Waterloo, Canada, January 2007.

- [7] 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.

- [8] 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, August 2006.

- [9] P. Emeliyanenko. Visualization of points and segments of real algebraic plane curves. Masters thesis, Universität des Saarlandes, February 2007.

[10] L. Gonzalez-Vega and I. Necula. Efficient topology determination of implicitly defined algebraic plane curves. *Computer Aided Geometric Design*, 19:719–743, 2002.

- [11] M. Kerber. Analysis of real algebraic plane curves. Masters thesis, Universität des Saarlandes, September 2006.

- [12] M. Kerber. Division-free computation of subresultants using bezout matrices. Research Report MPI-I-2006-1-006, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, May 2006.

- [13] L. Kettner. Reference counting in library design—optionally and with union-find optimization. In D. Musser and J. Siek, eds., *Proceedings of the First International Workshop on Library-Centric Software Design, LCSD'05*, San Diego, CA, USA, 2006, *Technical Report*, vol. 06-12, pp. 34–43. Rensselaer Polytechnic Institute, Computer Science Department.

[14] S. Lazard, L. M. Peñaranda, and S. Petitjean. Intersecting quadrics: an efficient and exact implementation. In *20th ACM Symposium on Computational Geometry (SCG'04)*, 2004, pp. 419–428.

[15] M. Meyerovitch. Robust, generic and efficient construction of envelopes of surfaces in three-dimensional spaces. In *ESA*, 2006, pp. 792–803.

[16] B. Mourrain, J.-P. Técourt, and M. Teillaud. On the computation of an arrangement of quadrics in 3d. *Comput. Geom. Theory Appl.*, 30(2):145–164, 2005.

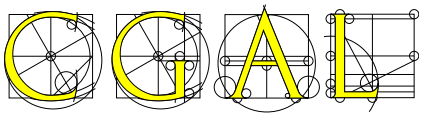
• [17] R. Seidel and N. Wolpert. On the exact computation of the topology of real algebraic curves. In J. S. B. Mitchell and G. Rote, eds., *Proceedings of the 21st ACM Symposium on Computational Geometry*, Pisa, Italy, 2005, pp. 107–115. ACM.

[18] A. Strzebonski. Cylindrical algebraic decomposition using validated numerics. *Journal of Symbolic Computation*, 41:1021–1038, 2006.

[19] R. Wein, E. Fogel, B. Zukerman, and D. Halperin. Advanced programming techniques applied to CGAL's arrangement package. In *1st Wrkshp. Library-Centric Software Design. (LCSD'06)*, 2005.

[20] C. K. Yap. Towards exact geometric computation. *CGTA: Computational Geometry: Theory and Applications*, 7, 1997.

12.8.2 CGAL: Computational Geometry Algorithms Library



We have continued the development of CGAL, the *Computational Geometry Algorithms Library*[8]. The distinguishing features of the library are the careful and efficient treatment of robustness issues, the wide scope of the algorithms and data structures provided, flexibility, extensibility, and ease of use. There have been two releases of the library since 2005 (3.2 in May 2006 and 3.2.1 in July 2006). The next release 3.3 is scheduled for April 2007.

Development and Maintenance

Investigators: Eric Berberich, Michael Hemmer, Peter Hachenberger, Lutz Kettner, Sebastian Limbach, and Andreas Meyer

A major effort in CGAL is the maintenance and improvement of the existing code base, user support, and the creation of new releases. Several people at MPI contribute here. We give an overview of tasks done at MPI:

- Implementation of three-dimensional Nef polyhedra, see below.
- Maintaining the following packages at MPI: Nef_S2, Nef_3, Box_intersection_d, Manual_tools
- Maintenance of test-suite, web page and CGAL mailing lists
- In the first year of the reporting period Lutz Kettner was an active member of the CGAL Editorial Committee and reviewed new submissions to CGAL.

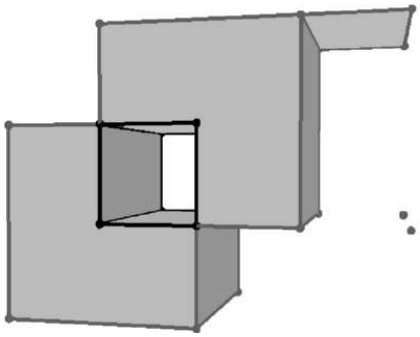
[8] `http://www.cgal.org/`

Nef Polyhedra in 3D

Investigators: Miguel Granados, Peter Hachenberger, Susan Hert, Lutz Kettner, Kurt Mehlhorn, and Michael Seel

A Nef polyhedron is any set that can be obtained from open half-spaces by a finite number of set complement and set intersection operations [4, 1]. The set of Nef polyhedra is closed under the Boolean set operations.

We implemented a data structure, the Selective Nef Complex, that realizes three-dimensional Nef polyhedra. Our implementation supports the construction of Nef polyhedra from manifold solids, boolean operations (union, intersection, complement, difference, symmetric difference), topological operations (interior, closure, boundary, regularization), and rational transformations including rotations by rational rotation matrices. Our software module is part of CGAL.



Our implementation is exact. We follow the exact computation paradigm to guarantee correctness. The binary operations are realized by a complete but currently inefficient algorithm for the overlay of two Nef polyhedra. The algorithm is complete in the sense that it can handle all inputs and requires no general position assumption.

At the end of the previous period, the functionality of our package was complete, robust, and optimized for efficiency [3]. As an application of Nef polyhedra, we now realized the Minkowski sum of closed three-dimensional polyhedra [2].

The Minkowski sum of two point sets P and Q in Euclidean space is the set of points $\{p + q : p \in P, q \in Q\}$. Minkowski sums are often used in motion planning for non-trivially shaped robots with translational movement. Here, the Minkowski sum of the robot and the inverse of a set of obstacles describes the set of all illegal robot placements. Correctness is crucial for this problem, especially for the handling of tight passages, i.e., passages through the obstacles that are exactly as wide as the robot (see Figure 12.14).

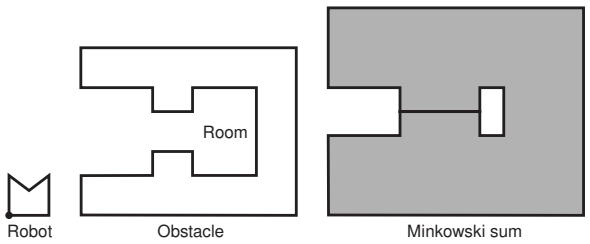


Figure 12.14: The Minkowski sum of a translational robot and an obstacle. There is only a single path (tight passage) that allows the robot to enter/leave the room.

Our implementation decomposes each polyhedron into convex sub-polyhedra, computes all pairwise Minkowski sums of the sub-polyhedra, and finally unites the pairwise sums. Currently, all other implementations rely on approximation schemes. Our implementation

is the only one that correctly computes the Minkowski sum of two closed three-dimensional polyhedra. As the next step we want to allow open polyhedra, which is the final step needed to solving tight passage problems.

References

[1] H. Bieri. Nef polyhedra: A brief introduct. *Comp. Suppl. Springer Verlag*, 10:43–60, 1995.

- [2] P. Hachenberger. *Boolean Operations on 3D Selevctive Nef Complexes: Data Structure, Algorithms Optimized Implementation, Experiments and Applications*. Phd thesis, Universität des Saarlandes, December 2006.
- [3] 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, June 2005, pp. 163–174. ACM.

[4] W. Nef. *Beiträge zur Theorie der Polyeder*. Herbert Lang, Bern, 1978.

12.8.3 Controlled Perturbation

Investigators: Manuel Caroli, Kurt Mehlhorn, Andreas Meyer, Ralf Osbild, and Michael Sagraloff

New algorithms in computational geometry are often designed assuming exact real arithmetic and non-degenerate input. Their straight forward implementation using floating point arithmetic usually leads to wrong results and robustness problems for some inputs due to rounding errors. Kettner et al. [5] had shown what can go wrong even with algorithms as simple as a planar convex hull computation. One way to overcome the robustness problem is the usage of exact arithmetic that naturally consumes more running time, and leave it to the programmer to handle *all* degenerate inputs which are barely considered in theoretical publications.

Controlled Perturbation is a framework that allows the implementation of such "idealistic algorithms" without the need to explicitly handle those problems. In other words it allows to write a *simple* but *fast* implementation, that computes the *correct* result to a slightly perturbed input. Controlled perturbation was pioneered by Halperin et al. and applied by them to the arrangement computation for spheres [4], polyhedral surfaces [7] and circles [3]. Funke et al. [2] had shown that controlled perturbation is a reasonable conversion tool for most randomized incremental algorithms. They further derived concrete controlled perturbation schemes for the construction of Delaunay triangulations and convex hulls.

In [6] we showed that the idea of controlled perturbation can be applied to a much wider class of geometric algorithms. Further we gave a general approach to analyze such algorithms. In [1] we applied this approach to more and less common predicates.

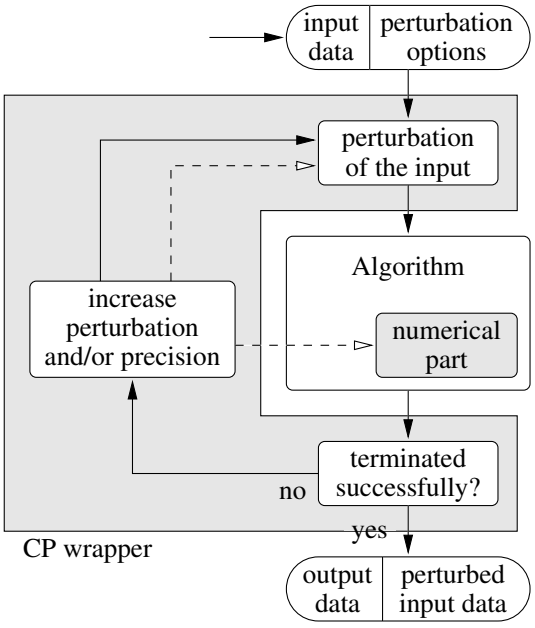
Controlled perturbation *numerically* perturbs the input randomly and independently in such a way that degeneracies disappear and all computations can be handled with fixed precision floating point arithmetic. Due to rounding errors we therefore have to avoid perturbations that lead to values too close to zero and we call this loci *forbidden region*. When devising a concrete controlled perturbation scheme, one is interested in the precision needed to resolve all problems for a fixed given perturbation amount, or, in the perturbation amount

needed for given precision. To analyze this, we estimate the number and size of the forbidden regions for an input object to be inserted. Note that by construction the measures of forbidden regions depend on the chosen precision. Perturbation amount and precision then have to be adapted in such a way that the current object will not lie inside any of those forbidden regions with high probability.

We argue that controlled perturbation can be applied to all algorithms as long as we can *a priori* upper bound the maximum error for *every* predicate evaluation that is responsible for a decision. This is the case for many predicates, e. g. for polynomials in input values.

Our general approach using separation explains for every predicate how to relate numerically the precision of the floating point arithmetic to the amount of perturbation while retaining the success probability. Because of its generality, however, the derived inequalities are often not tight and can be further improved manually for certain predicates. We examined predicates used during the computation of a circle arrangement and Voronoi diagram of line segments.

Following the idea of applying controlled perturbation to a wide class of algorithms, we developed a prototype in the context of the CGAL library. It can be applied to (already existing) robust implementations as well as to (new) implementations that don't take care of degeneracies. The controlled perturbation mechanism consists of a wrapper that perturbs the input and repeatedly calls the algorithm in question until the computation succeeds. To force this, the wrapper increases the amount of perturbation or the precision with each iteration. To detect if the computation succeeds, i. e. the values of predicates are sufficiently far from zero, we use a thin layer between predicates and algorithms that is completely independent of both. Currently it uses interval arithmetic with arbitrary precision floating point. This way, we always compute efficiently with the lowest possible precision at that time, regardless of the theoretical worst-case estimate.



References

- [1] M. Caroli. Evaluation of a generic method for analyzing controlled-perturbation algorithms. Masters thesis, Universität des Saarlandes, March 2007.
- [2] S. Funke, C. Klein, K. Mehlhorn, and S. Schmitt. Controlled perturbation for delaunay triangulations. In *Proceedings of the sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-05)*, Vancouver, Canada, 2005, pp. 1047–1056. SIAM.

[3] D. Halperin and E. Leiserowitz. Controlled perturbation for arrangements of circles. *International*

Journal of Computational Geometry & Applications, 14(4 & 5):277–310, 2004. Special issue, papers from SoCG 2003.

[4] D. Halperin and C. R. Shelton. A perturbation scheme for spherical arrangements with application to molecular modeling. *Comput. Geom. Theory Appl.*, 10:273–287, 1998.

[5] L. Kettner, K. Mehlhorn, S. Pion, S. Schirra, and C. Yap. Classroom examples of robustness problems in geometric computations. In S. Albers and T. Radzik, eds., *ESA 2004: 12th Annual European Symposium on Algorithms*, Bergen, Norway, September 2004, *LNCS 3221*, pp. 702–713. Springer.

- [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] S. Raab. Controlled perturbation for arrangements of polyhedral surfaces with application to swept volumes. In *Proc. 15th ACM Symposium on Computational Geometry*, 1999, pp. 163–172.

12.9 Academic Activities

12.9.1 Journal Positions

Kurt Mehlhorn is on the editorial board of

- *Computational Geometry Theory and Applications* (Editor in Chief),
- *ACM Transactions on Algorithms.*

12.9.2 Conference and Workshop Positions

Membership in Program Committees

Holger Bast:

- *3rd Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN)* , United Kingdom, July 2006,
- *29th Annual Conference on Research and Development on Information Retrieval (SIGIR)*, Seattle, August 2006,
- *29th European Conference on Information Retrieval (ECIR)*, Rome, April 2007,
- *30th Annual Conference on Research and Development on Information Retrieval (SIGIR)*, Amsterdam, July 2007.

Benjamin Doerr:

- *17th International Symposium on Algorithms and Computation (ISAAC 2006)*, Kolkata, December 2006.
- Mini-symposium “Hypergraphs” at the DMV Jahrestagung 2006, Bonn, September 2006 (Organizer).

Naveen Garg:

- *26th Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, Kolkata, December 2006 (Program Chair).

Joachim Giesen:

- *23rd Annual ACM Symposium on Computational Geometry (SOCG)*, Gyeongju, June 2007,
- *24th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Aachen, February 2007,
- *3d Symposium on Point-Based Graphics*, Boston, July 2006.

Kurt Mehlhorn:

- *6th Workshop on Experimental Algorithms (WEA)*, Rome, June 2007,
- *European Symposium on Algorithms*, 2008 (Program Chair).

Ulrich Meyer:

- *International Conference on High Performance Computing (HiPC)*, Goa, December 2007,
- *34th International Colloquium on Automata, Languages and Programming (ICALP), Track A*, Wroclaw, July 2007,
- *Genetic and Evolutionary Computation Conference (GECCO 2007)*, London, July 2007,
- *33rd International Conference on Current Trends in Theory and Practice of Computer Science(SOFSEM)*, Harrachov, January 2007,
- *14th Annual European Symposium on Algorithms (ESA), Track B*, Zürich, September 2006,
- *5th International Workshop on Experimental Algorithms (WEA)*, Menorca Island, May 2006,
- *IEEE International Conference on Distributed Computing Systems (ICDCS)*, Lisboa, July 2006,
- *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Rhodes Island, April 2006.

Seth Pettie:

- *18th Symposium on Discrete Algorithms (SODA)*, New Orleans, January 2007.

René Sitters:

- *8th Workshop on Models and Algorithms for Planning and Scheduling Problems (MAPSP)*, Istanbul, July 2007.

12.9.3 Invited Talks and Tutorials

Holger Bast:

- *Why Spectral Retrieval Works (and when it does not)*, Dagstuhl Seminar: Algorithmic Aspects of Large and Complex Networks, September 7, 2005.
- *Type Less, Find More: Fast Autocompletion Search with a Succinct Index*, Google Inc., Mountain View, USA, August 14, 2006.
- *The CompleteSearch Engine*, Yahoo! Research, Santa Clara, USA, January 5, 2007,
- *Why Spectral Retrieval Works (and when it does not)*, Dagstuhl Seminar: Web Information Retrieval and Linear Algebra Algorithms, February 14, 2007.
- *The CompleteSearch Engine*, The Future of Web Search, Bertinoro Italy, June 19, 2007.
- *Transit: Ultrafast Shortest-Path Queries on Road Networks*, Algorithms and Data Structures, Bertinoro Italy, October 3, 2007.

Khaled Elbassioni

- *Conflict-Free Colorings of Rectangular Ranges*, Invited talk, Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, May 2006.

Stefan Funke:

- *Algorithmische Geometrie – von der Theorie zur Anwendung und zurück*, Plenary talk, 3.Tagung der Deutschen Gesellschaft für Geometrie und Grafik (DGfGG), Hochschule für Technik, Stuttgart, June 2007

Joachim Giesen:

- *Conformal Alpha Shapes*, Invited talk, Workshop on Describing Multivariate Distributions with Nonlinear Variation, Radcliffe Institute for Advanced Study, Harvard University, Cambridge, USA, October 2006.

Kurt Mehlhorn:

- European Symposium on Algorithms (ESA), Plenary talk, Zürich, 2006.
- German Conference on Operations Research, Plenary talk, Karlsruhe, 2006.
- Italian Theory Conference, Invited talk, Rome, 2006.
- Dutch Theory Day, Invited talk, Utrecht, 2006.
- Aarhaus University Distinguished Speaker Series, 2006.
- International Symposium on Mathematical Foundations of Computer Science (MFCS), Invited talk, 2007.

12.9.4 Other Academic Activities

- ADFOCS organized by Benjamin Doerr and Michael Sagraloff, Saarbrücken, August 2006.
- Kurt Mehlhorn has served as Vice President of MPG since 2002 (until 2008).
- Kurt Mehlhorn was on the international review panel of British ICT (December 2006).

12.10 Teaching Activities

Summer Semester 2005

Courses:
Algorithm Engineering, (L. Kettner, K. Mehlhorn and U. Meyer)
Optimization, (E. Althaus and B. Doerr)
Online and Approximation Algorithms, (U. Meyer)

Seminars:
Geometrische Algorithmen, (S. Schmitt and N. Wolpert)

Winter Semester 2005/2006

Courses:
Data Structures and Algorithms, (M. Kutz and U. Meyer)
Approximation Algorithms, (K. Elbassioni and R. Sitters)
Discrepancy Theory, (B. Doerr)
Randomized Algorithms, (R. Beier, S. Canzar and S. Funke)
Datenstrukturen und effiziente Algorithmen, (E. Althaus–at Mainz University)

Seminars:
Advanced Index Data Structures, (H. Bast)
Computational and Algebraic Geometry, (K. Mehlhorn and F.O. Schreyer)
Bioinformatik, (E. Althaus–at Mainz University)

Summer Semester 2006

Courses:
Algorithms for Large Data Sets, (U. Meyer)
Computational Geometry, (S. Funke and J. Giesen)
Advanced Data Structures, (M. Mucha and S. Pettie)
Optimization, (R. Beier and B. Manthey)
Combinatorial Geometry, (S. Govindarajan and N. Mustafa)
Einführung in die Informatik für Hörer aller Fakultäten, (J. Siekmann, N. Hebbinghaus and C. Zinn)
Graphenalgorithmen, (E. Althaus–at Mainz University)

Theoretische Grundlagen der Informatik I,(E. Althaus–at Mainz University)

Seminars:
Liar Games and Noisy Channels, (B. Doerr and C. Klein)

Winter Semester 2006/2007

Courses:
Evolutionary Algorithms, (N. Hebbinghaus and F. Neumann)
Graph Theory, (B. Doerr)
Lineare Optimierung, (E. Althaus–at Mainz University)
Lecture Theoretische Grundlagen der Informatik II, (E. Althaus–at Mainz University)

Seminars:
Computational Topology (J. Giesen and M. Sagraloff)
Searching with Suffix Arrays (H. Bast)
New Trends on the Web, (J. Giesen and Ingmar Weber)
Topics in Algorithm Design, (U. Meyer–at Frankfurt University)
Bioinformatik, (E. Althaus–at Mainz University)

Bachelor Theses

- Manuel Caroli: *Exakte Arrangement-Berechnung gedrehter Quadratischer Kurven*, April 2006.
- Franziska Ebert: *Benchmark Data Sets for Conic Arrangements*, November 2005.
- Mahmoud Fouz: *Hereditary Discrepancies in Different Numbers of Colors*, October 2006.
- Daniel Schmitt: *Implementierung einer Überlagerung von konvexen Arrangements der Kugeloberfläche*, August 2006.
- David Steurer: *Tight Bounds on the Min-Max Boundary Decomposition Cost of Weighted Graphs*, May 2006.

Diploma Theses

- Deepak Ajwani: *Design, Implementation and Experimental Study of External Memory (BFS) Algorithms*, March 2005.
- Tim Aubertin: *Enhanced Proximity Search for the CompleteSearch Engine*, August 2006.
- Muhammad Kamran Azam: *Branch-and-Cut Techniques for Generalized Asymmetric Traveling Salesman Problem*, September 2005.
- Manuel Caroli: *Evaluation of a generic method for analyzing controlled-perturbation algorithms*, March 2007.
- Valentin Deleplace: *Exploring the Seeding-technique for Aligning Protein Sequences*, July 2005.

- Daniel Dumitriu: *Avoiding Slivers–In Practice and with a Guarantee*, April 2007.
- Pavel Emeliyanenko: *Visualization of Points and Segments of Real Algebraic Plane Curves*, February 2007.
- Daniel Fischer: *ALWIS: A Visualization Tool for Concept-Based Retrieval Schemes – Design and Implementation*, March 2006.
- Tobias Friedrich, *Deterministic Random Walks on Infinite Grids*, December 2005.
- Benedikt Grundmann: *ALWIS: A Visualization Tool for Concept-Based Retrieval Schemes – Theoretical Foundations and Models*, March 2006.
- Regis Newo Kenmogne: *Understanding LSI via the Truncated Term-Term Matrix*, May 2005.
- Michael Kerber: *Analysis of Real Algebraic Plane Curves*, September 2006.
- Jens Maue: *A Goal-Directed Shortest Path Algorithm Using Precomputed Cluster Distances*, June 2006.
- Sebastian Pohl: *Exact Integer Linear Programming with Bounded Variables in a Branch-and-Cut Algorithm*, March 2006.
- Daniel Schmitt: *WNETS–A Framework for Testing and Evaluating Algorithms and Models for Wireless Sensor Networks*, January 2007.
- Dominik Schultes: *Fast and Exact Shortest Path Queries Using Highway Hierarchies*, August 2005.
- David Steurer: *An Asymptotic Approximation Scheme for Multigraph Edge Coloring*, August 2006.
- Dennis Weber: *Solving Large Sparse Linear Systems Exactly*, August 2006.
- Caroline Weinand: *Fill-in Reduction while Solving Large, Sparse Linear Systems with Graph Theoretical Methods*, June 2005.

12.11 Dissertations, Habilitations, Offers, Awards

12.11.1 Dissertations

Completed:

- Roman Dementiev: *Algorithm Engineering for Large Data Sets*, December 2006.
- Dieter Mitsche: (ETH Zürich), *Spectral Methods for Reconstruction Problems*, 2006.
- Peter Hachenberger: *Boolean Operations on 3D Selective Nef Complexes: Data Structure, Algorithms Optimized Implementation, Experiments and Applications*, December 2006.
- Joachim Reichel: *Combinatorial Approaches to Trunk Packing*, July 2006.
- Dimitris Michail: *Minimum Cycle Basis, Algorithms and Applications*, July 2006.

In Preparation:

- Annamária Kovács: *Fast Algorithms for Two Scheduling Problems*, 2007.

- Debapriyo Majumdar: *On Spectral Retrieval and Efficient Top-k Query Processing*, March 2007.
- Ingmar Weber: *The CompleteSearch Engine*, June 2007.
- Domagoj Matijevic: *Geometric Optimization and Querying: Exact and Approximate*, 2007.
- Kanela Kagliosi, 2007.

12.11.2 Habilitations

Completed:

- Ernst Althaus, August 2006.
- Benjamin Doerr, *Integer Approximation.*
- Lutz Kettner, April 2006.
- Uli Meyer.

Ongoing:

- Holger Bast, June 2007.
- Stefan Funke, June 2007.

12.11.3 Offers for Faculty Positions

- Nicola Wolpert, FH Stuttgart, W2-Prof.
- Uli Meyer, Universität Frankfurt, W3 Prof.
- Seth Pettie, University of Michigan.
- Nabil Mustafa, Lahore University of Management Sciences.
- Surender Baswana, Indian Inst. of Tech. Kanpur.

12.11.4 Awards

- *Kurt Mehlhorn* received an honorary doctorate from the University of Waterloo in June 2006.
- *Naveen Garg* received an IBM faculty award for 2006.
- The paper "Packing a Trunk–now with a Twist," by *F. Eisenbrand, S. Funke, A. Karrenbauer, J. Reichel and E. Schmömer*, won the second best paper award at the ACM Symposium on Solid and Physical Modeling (SPM) in 2005.
- For his Diploma Thesis, entitled "Deterministic Random Walks on Infinite Grids", supervised by Benjamin Doerr, *Tobias Friedrich* received an *Examenspreis 2006* from the Faculty of Mathematics and Computer Science at the University of Jena.
- *Arno Eigenwillig* received an ISSAC 2006 Distinguished Student Author award.

12.12 Grants and Cooperations

We participate in some EU-projects (DELIS, APPOL II, and ACS), a GIF project, a DFG-project, and one industry funded project. Recently a scientific partner group was installed at IIT Delhi 12.4.3. There are also many smaller cooperations which do not receive extra funding and are not listed here.

12.12.1 Projects Funded by the European Union

Dynamically Evolving Large-Scale Information Systems (DELIS)

DELIS, which stands for "Dynamically Evolving Large Scale-Information Systems", is a huge integrated European project under the roof of the 6th Framework Programme. DELIS comes in six so-called subprojects (SPs): "System's Monitoring", "Self-Organization", "Large-Scale Optimization", "Game Theory", "Biologically-Inspired Computing", "Peer-to-Peer Web Search" (titles abridged). The goal of the project is to bring together, in a big interdisciplinary effort and as part of the *Complex Systems* initiative, European researchers from diverse communities such as computer science, economy, physics, biology, and industry. There are twenty partners, many of which were already partners in the previous ALCOM projects; the new partners are mainly from industry or areas other than computer science. For details, see the DELIS website at `http://delis.upb.de`.

The project started in January 2004 and was just recently approved for continuation; the whole project duration is 3 years. The institute is involved mainly in SP3 (Large-Scale Optimization) and SP6 (Peer-to-Peer Web Search). A number of former staff members or postdocs are now partners in DELIS. Currently, one of the institute's postdocs is payed from DELIS money. A former PhD student from D1 is payed as a postdoc from DELIS money in Rome.

In the first year of DELIS, D1 has contributed work in combinatorial optimization (12.4), external-memory computation (Section 12.6.1), and information retrieval (Section 12.7). Holger Bast coordinated two major deliverables of SP3. For the contributions of SP6, see Section 16.11.

ACS

The project "Algorithms for Complex Shapes", is a joint effort together with six other European research groups and it aims to advance the handling of non linear objects in geometric algorithms, both in theory and practice. This requires interdisciplinary cooperation of computational geometry with computer algebra and numerical analysis, as well as with software engineering and computer aided design. The EU grant support began on March 1, 2005 and runs for 36 months. The partners and group leaders of the ACS project are:

- *INRIA Sophia-Antipolis*, France (Dr. J.-D. Boissonnat, Dr. B. Mourrain, Dr. M. Teillaud)
- *ETH, Zürich,* Switzerland, (B. Gärtner, Prof. E. Welzl, Prof. P. Widmayer)
- *Freie Universität Berlin,* Germany (Prof. H. Alt, Prof. G. Rote)
- *Rijksuniversiteit Groningen,* The Netherlands (Prof. G. Vegter)

- *Max Planck Institute for Informatics,* Saarbrücken, Germany (Prof. K. Mehlhorn, Dr. M. Sagraloff)
- *Tel Aviv University,* Israel (Prof. D. Halperin)

12.12.2 Graph Algorithms: Theory and Practice (funded by GIF)

The project *Graph Algorithms: Theory and Practice* is funded by the German-Israeli Foundation (GIF) for Scientific Research and Development. It was started in January 2005 and will run for three years. It is led by Kurt Mehlhorn at MPI and Uri Zwick at Tel-Aviv University. Under the project number I-792-136.6 GIF essentially covers the salaries of two PhD students, one at each site.

We are planning to consider various aspects of graph problems. We will treat both *static* and *dynamic* versions of several problems. In addition to the standard *worst-case* complexity we will also study the *average case* complexity of some of the problems, and the recently introduced *smoothed complexity* (Spielman and Teng). We will also attempt to produce algorithms that *certify* their results, i.e., algorithms that produce a succinct, easily verifiable, proof that the result they produced is correct. Finally, we will also try to examine various tradeoffs possible between the running time and the *accuracy* of the solution obtained.

12.12.3 Cooperations With Industry

Algorithmic Solutions (AS) is a spin-off of D1. AS is marketing LEDA (Library of Efficient Algorithms) under license agreements with Max Planck Society. It is also responsible for the maintenance. Through its industrial contacts and through user feedback, AS brings interesting research problems into D1. CGAL is marketed by Geometry Factory, a spin-off of our EU-projects CGAL and GALIA.

With Daimler-Chrysler we are cooperating on assembly simulation and geometric packing. Joachim Reichel is paid by this cooperation. The information retrieval group established a cooperation with Recommind Inc.

12.12.4 Partner Group Approximation Algorithms at IIT Delhi (funded by MPG)

Scientific Partner Groups of Max Planck Society (Max Planck Partner Groups) can be established together with a foreign research institute when an outstanding junior scientist, subsequent to completing his or her research residency at a Max Planck Institute, returns to a productive laboratory in his homeland in order to continue research work that is also in the interest of his or her former hosting Max Planck Institute. Lasting relations between the Max Planck Institutes and former foreign guest scientists are the goal of these Partner Groups. To this end, Max Planck Society makes 20,000 euros per Partner Group available over a five-year period.

Naveen Garg, a former member of D1, leads one of the first four Max Planck Partner Groups, which were inaugurated in December 2004. Naveen Garg is a computer scientist at IIT Delhi to which he returned in 2000 after several years of research residency at MPI Saarbrücken. Since then, Naveen Garg has been working closely together with Kurt

Mehlhorn's group. Garg's research interest lies in the area of approximation algorithms and combinatorial optimization. His group focuses on flow and network optimization problems. See Section 12.4.3.

12.13 Publications

Books

[1] J. Mitchell, G. Rote, and L. Kettner, eds. *21st Annual Symposium on Computational Geometry (SCG'05)*, New York, USA, June 2005. ACM.

Journal articles and book chapters

[1] D. J. Abraham, K. Cechlárová, D. Manlove, and K. Mehlhorn. Pareto optimality in house allocation problems. In X. Deng and D. Du, eds., *Algorithms and computation, 16th International Symposium, ISAAC 2005*, Sanya, Hainan, China, 2005, *LNCS 3827*, pp. 1163–1175. Springer.

[2] E. Althaus, S. Funke, S. Har-Peled, J. Könemann, E. A. Ramos, and M. Skutella. Approximating k-hop minimum-spanning trees. *Operations Research Letters*, 33(2):115–120, March 2005.

[3] I. Bárány and B. Doerr. Balanced partitions of vector sequences. *Linear Algebra and Its Applications*, 414(2-3):464–469, 2006.

[4] H. Bast. Präzisionssuche im homöopathischen Netz. *Homöopathie-Zeitschrift*, I/05:119–122, 2005.

[5] H. Bast. Intelligente Suche mit garantiert schnellen Antwortzeiten. *MPG Jahrbuch*, 2006.

[6] H. Bast, K. Mehlhorn, G. Schäfer, and H. Tamaki. Matching algorithms are fast in sparse random graphs. *Theory of Computing Systems*, 39(1):3–14, February 2006.

[7] R. Beier, A. Czumaj, P. Krysta, and B. Vöcking. Computing equilibria for a service provider game with (im)perfect information. *ACM Transactions on Algorithms*, 2(4):679–706, October 2006.

[8] R. Beier and B. Vöcking. An experimental study of random knapsack problems. *Algorithmica*, 45(1):121–136, February 2006.

[9] R. Beier and B. Vöcking. Typical properties of winners and losers in discrete optimization. *SIAM Journal on Computing*, 35(4):855–881, February 2006.

[10] N. Beldiceanu, I. Katriel, and S. Thiel. Filtering algorithms for the same and usedby constraints. *Archives of Control Sciences*, 2007. to appear.

[11] H. Brönnimann, L. Kettner, M. Pocchiola, and J. Snoeyink. Counting and enumerating pointed pseudotriangulations with the greedy flip algorithm. *SIAM Journal on Computing*, 36(3):721–739, October 2006.

[12] D. Bundy, N. Hebbinghaus, and B. Stellmacher. The local c(g,t) theorem. *Journal of Algebra*, 300(2):741–789, June 2006.

[13] M. Chrobak, C. Dürr, W. Jawor, L. Kowalik, and M. Kurowski. A note on scheduling equal-length jobs to maximize throughput. *Journal of Scheduling*, 9(1):71–73, 2006.

[14] R. Cole and L. Kowalik. New linear-time algorithms for edge-coloring planar graphs. *Algorithmica*, 2007. To appear.

[15] R. Cole, L. Kowalik, and R. Skrekovski. A generalization of Kotzig's theorem and its application. *SIAM Journal on Discrete Mathematics*, 21:93–106, 2007.

[16] A. Dessmark, P. Fraigniaud, D. Kowalski, and A. Pelc. Deterministic rendezvous in graphs. *Algorithmica*, 46(1):69–96, 2006. Part of this work was published in ISAAC 2004.

[17] B. Doerr. Matrix rounding with respect to small submatrices. *Random Structures & Algorithms*, 28(1):107–112, 2006.

[18] B. Doerr. Non-independent randomized rounding and coloring. *Discrete Applied Mathematics*, 154(4):650–659, 2006.

[19] B. Doerr. Lp linear discrepancy of totally unimodular matrices. *Linear Algebra and Its Applications*, 420:663–666, 2007.

[20] B. Doerr. Matrix approximation and tusnady's problem. *European Journal of Combinatorics*, 28:990–995, 2007.

[21] B. Doerr and T. Friedrich. Quasirandomness in graphs. *Electronic Notes in Discrete Mathematics*, 25:61–64, 2007.

[22] B. Doerr, M. Gnewuch, and N. Hebbinghaus. Discrepancy of symmetric products of hypergraphs. *The Electronic Journal of Combinatorics*, 13:1–12, 2006.

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

[24] B. Doerr, N. Hebbinghaus, and F. Neumann. Speeding up evolutionary algorithms through unsymmetric mutation operators. *Evolutionary Computation*, 2007. To appear.

[25] B. Doerr, N. Hebbinghaus, and S. Werth. Improved bounds and schemes for the declustering problem. *Theoretical Computer Science*, 359(1-3):123–132, 2006.

[26] B. Doerr and C. Klein. Controlled randomized rounding. *Electronic Notes in Discrete Mathematics*, 25:39–40, 2006.

[27] B. Doerr and U. Lorenz. Error propagation in game trees. *Mathematical Methods of Operations Research*, 64(1):79–93, 2006.

[28] D. Donato, L. Laura, S. Leonardi, U. Meyer, S. Millozzi, and J. F. Sibeyn. Algorithms and experiments for the webgraph. *Journal of Graph Algorithms and Applications*, 10(2), 2007.

[29] A. Eigenwillig. On multiple roots in Descartes' rule and their distance to roots of higher derivatives. *Journal of Computational and Applied Mathematics*, 200(1):226–230, March 2007.

[30] 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, August 2006.

[31] K. Elbassioni, A. Elmasry, and I. Kamel. An indexing method for answering queries on moving objects. *Distributed and Parallel Databases*, 17(3):215–249, 2005.

[32] K. Elbassioni, Z. Lotker, and R. Seidel. Upper bound on the number of vertices of polyhedra with 0, 1-constraint matrices. *Information Processing Letters*, 100(2):69 – 71, October 2006.

[33] C. Feremans, A. Grigoriev, and R. Sitters. The geometric generalized minimum spanning tree problem with grid clustering. *4OR: A Quarterly Journal of Operations Research*, 4(4):319–329, 2006.

[34] E. Fogel, D. Halperin, L. Kettner, M. Teillaud, R. Wein, and N. Wolpert. Arrangements. In J.-D. Boissonnat and M. Teillaud, eds., *Effective Computational Geometry for Curves and Surfaces*, Mathematics and Visualization, ch. 1, pp. 1–66. Springer, Berlin, Germany, 2007.

[35] D. Fotakis, R. Pagh, P. Sanders, and P. G. Spirakis. Space efficient hash tables with worst case constant access time. *Theory of Computing Systems*, 38(2):229–248, February 2005.

[36] S. Funke, A. Kesselman, U. Meyer, and M. Segal. A simple improved distributed algorithm for minimum CDS in unit disk graphs. *ACM Transactions on Sensor Networks*, 2(3):444–453, 2006.

[37] S. Funke, K. Mehlhorn, and S. Näher. Structural filtering: a paradigm for efficient and exact geometric programs. *Computational Geometry*, 31(3):179–194, 2005.

[38] C. Georgiou, D. Kowalski, and A. Shvartsman. Efficient gossip and robust distributed computation. *Theoretical Computer Science*, 347(1):130–166, 2005. submitted to journal. The extended abstract appeared DISC 2003.

[39] C. Gotsman, K. Kaligosi, K. Mehlhorn, D. Michail, and E. Pyrga. Cycle bases of graphs and sampled manifolds. *Computer Aided Geometric Design*, 2007. To appear.

[40] D. A. Hutchinson, P. Sanders, and J. S. Vitter. Duality between prefetching and queued writing with parallel disks. *SIAM Journal on Computing*, 34(6):1443–1463, 2005.

[41] R. W. Irving, T. Kavitha, K. Mehlhorn, D. Michail, and K. Paluch. Rank-maximal matchings. *ACM Transactions on Algorithms*, 2(4):602–610, October 2006.

[42] D. Johannsen. A direct decomposition of 3-connected planar graphs. *Séminaire Lotharingien de Combinatoire*, 54A:15 pp, 2007.

[43] I. Katriel, L. Michel, and P. Van Hentenryck. Maintaining longest paths incrementally. *Constraints*, 10(2):159–183, 2005.

[44] I. Katriel and S. Thiel. Complete bound consistency for the global cardinality constraint. *Constraints*, 10(3):191–217, 2005.

[45] A. Kesselmann, D. Kowalski, and M. Segal. Energy efficient communication in ad hoc networks from user's and designer's perspective. *ACM SIGMOBILE Mobile Computing and Communications Review*, 9(1):15–26, 2005.

[46] L. Khachiyan, E. Boros, K. Elbassioni, V. Gurvich, and K. Makino. On the complexity of some enumeration problems for matroids. *SIAM Journal on Discrete Mathematics*, 19(4):966–984, 2005.

[47] L. Kowalik. Adjacency queries in dynamic sparse graphs. *Inf. Proc. Lett.*, 2007.

[48] D. Kowalski and A. Pelc. Time complexity of radio broadcasting: adaptiveness vs. obliviousness and randomization vs. determinism. *Theoretical Computer Science*, 333(3):355–371, 2005.

[49] W. Krandick and K. Mehlhorn. New bounds for the descartes method. *Journal of Symbolic Computation*, 41(1):49–66, 2006.

[50] D. Kratsch, R. McConnell, K. Mehlhorn, and J. P. Spinrad. Certifying algorithms for recognizing interval graphs and permutation graphs. *SIAM Journal on Computing*, 36(2):326–353, 2006.

[51] M. Kutz. Conway's angel in three dimensions. *Theoretical Computer Science*, 349(3):443–451, December 2005.

[52] K. Mehlhorn and D. Michail. Implementing minimum cycle basis algorithms. *ACM Journal of Experimental Algorithmics*, 11:1–14, 2006.

[53] K. Mehlhorn and D. Michail. Implementing minimum cycle basis algorithms. *Journal of Experimental Algorithmics*, 11:1–14, 2007.

[54] B. Mourrain, S. Pion, S. Schmitt, J.-P. Técourt, E. Tsigaridas, and N. Wolpert. Algebraic issues in computational geometry. In J.-D. Boissonnat and M. Teillaud, eds., *Effective Computational Geometry for Curves and Surfaces*, Mathematics and Visualization, ch. 3, pp. 117–155. Springer, Berlin, Germany, 2007.

[55] F. Neumann. Expected runtimes of evolutionary algorithms for the eulerian cycle problem. *Computers and Operations Research*, 2007. To appear.

[56] F. Neumann and I. Wegener. Randomized local search, evolutionary algorithms, and the minimum spanning tree problem. *Theoretical Computer Science*, 2007. To appear.

[57] E. Schömer and N. Wolpert. An exact and efficient approach for computing a cell in an arrangement of quadrics. *Computational Geometry*, 33(1-2):65 – 97, 2006.

[58] R. Sitters. Complexity of preemptive minsum scheduling on unrelated parallel machines. *Journal of Algorithms*, 57(1):37–48, 2005.

[59] R. Sitters and L. Stougie. The general two-server problem. *Journal of the ACM*, 53(3):437–458, 2006.

[60] S. Teramoto, T. Asano, N. Katoh, and B. Doerr. Inserting points uniformly at every instance. *IEICE - Transactions on Information and Systems*, E89-D(8):2348–2356, 2006.

Conference articles

[1] D. J. Abraham, K. Cechlárová, D. Manlove, and K. Mehlhorn. Pareto optimality in house allocation problems. In X. Deng and D. Du, eds., *Algorithms and computation, 16th International Symposium, ISAAC 2005*, Sanya, Hainan, China, 2005, *LNCS 3827*, pp. 1163–1175. Springer.

[2] N. Ahuja, A. Baltz, B. Doerr, A. Privetivy, and A. Srivastav. On the minimum load coloring problem. In T. Erlebach and P. Persiano, eds., *Third Workshop on Approximation and Online Algorithms (WAOA 2005)*, Palma de Mallorca, Spain, 2005, *LNCS 3879*, pp. 15–26. Springer.

[3] D. Ajwani, R. Dementiev, and U. Meyer. A computational study of external memory BFS algorithms. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'06*, Miami, USA, 2006, pp. 601–610. ACM-SIAM.

[4] D. Ajwani, K. 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, June 2007, Proceedings. ACM. To Appear.

[5] D. Ajwani, T. Friedrich, and U. Meyer. An $o(n^{2.75})$ algorithm for online topological ordering. In L. Arge and R. Freivalds, eds., *Algorithm theory - SWAT 2006, 10th Scandinavian Workshop on Algorithm Theory*, Riga, Latvia, 2006, *LNCS 4059*, pp. 53–64. Springer.

[6] D. Ajwani, U. Meyer, and V. Osipov. Improved external memory BFS implementations. In D. Applegate and G. Brodal, eds., *9th Workshop on Algorithm Engineering and Experiments (ALENEX)*, New Orleans, USA, 2007. SIAM.

[7] 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.

[8] S. Arya, T. Malamatos, and D. M. Mount. The effect of corners on the complexity of approximate range searching. In N. Amenta and O. Cheong, eds., *Proceedings of the 22nd Annual Symposium on Computational Geometry, SCG'06*, Sedona, Arizona, USA, 2006, pp. 11–20. ACM.

[9] S. Arya, T. Malamatos, and D. M. Mount. On the importance of idempotence. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing, STOC'06*, Seattle, USA, 2006, pp. 564–573. ACM.

[10] Y. Bartal, S. Leonardi, G. Shallom, and R. Sitters. On the value of preemption in scheduling. In J. Diaz, K. Jansen, J. D. P. Rolim, and U. Zwick, eds., *9th Workshop on Approximation Algorithms for Combinatorial Optimization Problems, APPROX*, Barcelona, Spain, August 2006, *LNCS 4110*, pp. 39–48. Springer.

[11] H. Bast, G. Dupret, D. Majumdar, and B. Piwowarski. Discovering a term taxonomy from term similarities using principal component analysis. In M. Ackermann, B. Berendt, M. Grobelnik, A. Hotho, D. Mladenic, G. Semeraro, M. Spiliopoulou, G. Stumme, V. Svatek, and M. W. van Someren, eds., *Semantics, web and mining, Joint International Workshops, EWMF 2005 and KDO 2005*, Porto, Portugal, 2006, *LNCS 4289*, pp. 103–120. Springer.

[12] H. Bast, S. Funke, and D. Matijevic. Transit: Ultrafast shortest-path queries with linear-time preprocessing. In C. Demetrescu, A. Goldberg, and D. Johnson, eds., *9th DIMACS Implementation Challenge — Shortest Path*, Piscataway, New Jersey, 2006. DIMACS.

[13] H. Bast, S. Funke, D. Matijevic, P. Sanders, and D. Schultes. In transit to constant time shortest-path queries in road networks. In D. Applegate and G. Brodal, eds., *9th Workshop on Algorithm Enginneering and Experiments (ALENEX'07)*, New Orleans, USA, 2007. SIAM.

[14] H. Bast and D. Majumdar. Why spectral retrieval works. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, eds., *28th Annual International Conference on Research and Development in Information Retrieval (SIGIR'05)*, Salvador, Brazil, August 2005, pp. 11–18. ACM.

[15] H. Bast, D. Majumdar, R. Schenkel, M. Theobald, and G. Weikum. IO-top-k at trec 2006: Terabyte track. In *The Fifteenth Text Retrieval Conference Proceedings, TREC 2006*, Gaithersburg, Maryland, USA, November 2006, pp. 1–5. NIST.

[16] 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.

[17] H. Bast, C. W. Mortensen, and I. Weber. Output-sensitive autocompletion search. In F. Crestani, P. Ferragina, and M. Sanderson, eds., *String Processing and Information Retrieval, 13th International Conference, SPIRE 2006*, Glasgow, GB, 2006, *LNCS 4209*, pp. 150–162. Springer.

[18] H. Bast and I. Weber. Don't compare averages. In S. Nikoletseas, ed., *4th International Workshop on Efficient and Experimental Algorithms (WEA'05)*, Santorini Island, Greece, 2005, *LNCS 3503*, pp. 67–76. Springer.

[19] H. Bast and I. Weber. Insights from viewing ranked retrieval as rank aggregation. In J. Adachi, W. Shan, and A. Vakali, eds., *Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05)*, Tokyo, Japan, April 2005, pp. 243–248. IEEE.

[20] H. Bast and I. Weber. Type less, find more: Fast autocompletion search with a succinct index. In E. N. Efthimiadis, S. Dumais, D. Hawking, and K. Järvellin, eds., *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, USA, 2006, pp. 364–371. ACM.

[21] H. Bast and I. Weber. When you're lost for words: Faceted search with autocompletion. In A. Broder and Y. Maarek, eds., *SIGIR'06 Workshop on Faceted Search*, Seattle, USA, August 2006, pp. 31–35. ACM.

[22] H. Bast and I. Weber. The completesearch engine: Interactive, efficient, and towards ir & db integration. In G. Weikum, ed., *3rd Biennial Conference on Innovative Data Systems Research (CIDR'07)*, Asilomar, USA, 2007. VLDB Endowment.

[23] H. Bast and I. Weber. Managing helpdesk tasks with completesearch: A case study. In B. Decker and M. Nick, eds., *4th Conference on Professional Knowledge Management (WM'07)*, Berlin, Germany, 2007, LNCS. Springer.

[24] S. Baswana, V. Goyal, and S. Sen. All-pairs nearly 2-approximate shortest paths in $o(n^2\text{polylog}n)$ time. In V. Diekert and B. Durand, eds., *STACS 2005, 22nd Annual Symposium on Theoretical Aspects of Computer Science*, Stuttgart, Germany, 2005, *LNCS 3404*, pp. 666–679. Springer.

[25] R. Beier, H. Röglin, and B. Vöcking. The smoothed number of pareto optimal solutions in bicriteria integer optimization. In *12th Conference on Integer Programming and Combinatorial Optimization*, Ithaca, USA, 2007, pp. 000–999. Springer. To Appear.

[26] 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, March 2007, pp. 231–234. Technische Universitaet Graz.

[27] E. Berberich, A. Eigenwillig, M. Hemmer, S. Hert, L. Kettner, K. Mehlhorn, J. Reichel, S. Schmitt, E. Schömer, and N. Wolpert. Exacus: Efficient and exact algorithms for curves and surfaces. In G. S. Brodal and S. Leonardi, eds., *13th Annual European Symposium on Algorithms (ESA 2005)*, Palma de Mallorca, Spain, October 2005, *LNCS 3669*, pp. 155–166. Springer.

[28] 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, March 2007, pp. 223–226. Technische Universitaet Graz.

[29] E. Berberich, M. Hemmer, L. Kettner, E. Schömer, and N. Wolpert. An exact, complete and efficient implementation for computing planar maps of quadric intersection curves. In J. Mitchell, G. Rote, and L. Kettner, eds., *21st Annual Symposium on Computational Geometry (SCG'05)*, Pisa, Italy, June 2005, pp. 99–106. ACM.

[30] E. Berberich and M. Meyerovitch. Computing envelopes of quadrics. In *Proceedings of 23rd European Workshop on Computational Geometry*, Graz, Austria, March 2007, pp. 235–238. Technische Universitaet Graz.

[31] A. Bifet, C. Castillo, P.-E. Chirita, and I. Weber. An analysis of factors used in search engine ranking. In B. Davison, ed., *1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'05)*, Chiba, Japan, May 2005, pp. 1–10.

[32] H. L. Bodlaender, C. Feremans, A. Grigoriev, E. Penninkx, R. Sitters, and T. Wolle. On the minimum corridor connection and other generalized geometric problems. In T. Erlebach and C. Kaklamanis, eds., *4th Workshop on Approximation and Online Algorithms, WAOA*, Zurich, Switzerland, September 2006, *LNCS 4368*, pp. 69–82. Springer.

[33] 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 (GECCO-2007)*, London, UK, 2007. ACM. To appear.

[34] G. S. Brodal, K. Kaligosi, I. Katriel, and M. Kutz. Faster algorithms for computing longest common increasing subsequences. In L. Moshe and V. Gabriel, eds., *Combinatorial Pattern Matching, 17th Annual Symposium, CPM 2006*, Barcelona, Spain, 2006, *LNCS 4009*, pp. 330–341. Springer.

[35] H. Brönnimann, L. Kettner, M. Pocchiola, and J. Snoeyink. Counting and enumerating pointed pseudo-triangulations with the greedy flip algorithm. In C. Demetrescu, R. Tamassia, and R. Sedgewick, eds., *Proceedings of the Seventh Workshop on Algorithm Engineering and Experiments and the Second Workshop on Analytic Algorithmics and Combinatorics (ALENEX/ANALCO 2005)*, Vancouver, BC, Canada, January 2005, pp. 98–110. SIAM.

[36] K. L. Chang and R. Kannan. The space complexity of pass-efficient algorithms for clustering. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'06*, Miami, USA, January 2006, pp. 1157–1166. ACM-SIAM. This is the conference version. Full version under submission to journal.

[37] B. S. Chlebus and D. Kowalski. Cooperative asynchronous update of shared memory. In H. N. Gabow and R. Fagin, eds., *Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC 2005)*, Baltimore, USA, 2005, pp. 733–739. ACM.

[38] 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.

[39] J. Cooper, B. Doerr, J. Spencer, and G. Tardos. Deterministic random walks on the integers. In S. Felsner, ed., *2005 European Conference on Combinatorics, Graph Theory and Applications (EuroComb '05)*, Berlin, Germany, September 2005, *DMTCS Proceedings*, vol. AE, pp. 73–76. DMTCS.

[40] J. Cooper, B. Doerr, J. Spencer, and G. Tardos. Deterministic random walks. In R. Raman, R. Sedwgewick, and M. F. Stallmann, eds., *Proceedings of the Eighth Workshop on Algorithm Engineering and Experiments and the Third Workshop on Analytic Algorithmics and Combinatorics (ALENEX'06 / ANALCO'06)*, Miami, FL, USA, 2006, pp. 185–197. SIAM.

[41] J. Czyzowicz, D. Kowalski, E. Markou, and A. Pelc. Searching for a black hole in tree networks. In T. Higashino, ed., *Principles of distributed systems, 8th International Conference, OPODIS 2004*, Grenoble, France, 2005, *LNCS 3544*, pp. 67–80. Springer.

[42] R. Dementiev, L. Kettner, and P. Sanders. STXXL: Standard template library for XXL data sets. In G. S. Brodal and S. Leonardi, eds., *Algorithms - ESA 2005, 13th Annual European Symposium (ESA 2005)*, Palma de Mallorca, Spain, October 2005, *LNCS 3669*, pp. 640–651. Springer.

[43] 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.

[44] B. Doerr. Randomly rounding rationals with cardinality constraints and derandomizations. In W. Thomas and P. Weil, eds., *STACS 2007*, Aachen, 2007, *LNCS 4393*, pp. 441–452. Springer.

[45] 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, December 2006, *LNCS 4288*, pp. 474–483. Springer.

[46] B. Doerr, T. Friedrich, C. Klein, and R. Osbild. Rounding of sequences and matrices, with applications. In T. Erlebach and P. Persiano, eds., *Third Workshop on Approximation and Online Algorithms (WAOA 2005)*, Palma de Mallorca, Spain, 2005, *LNCS 3879*, pp. 96–109. Springer.

[47] B. Doerr, T. Friedrich, C. Klein, and R. Osbild. Unbiased matrix rounding. In L. Arge and R. Freivalds, eds., *Algorithm theory - SWAT 2006, 10th Scandinavian Workshop on Algorithm Theory*, Riga, Latvia, 2006, *LNCS 4059*, pp. 102–112. Springer.

[48] B. Doerr and N. Hebbinghaus. Discrepancy of products of hypergraphs. In S. Felsner, ed., *2005 European Conference on Combinatorics, Graph Theory and Applications (EuroComb '05)*, Berlin, Germany, September 2005, *DMTCS Proceedings*, vol. AE, pp. 323–328. DMTCS.

[49] B. Doerr, N. Hebbinghaus, and F. Neumann. Speeding up evolutionary algorithms through restricted mutation operators. In T. P. Runarsson, H. G. Beyer, E. Burke, J. J. Merelo-Guervós, L. D. Whitley, and X. Yao, eds., *Parallel Problem Solving from Nature - PPSN IX, 9th International Conference*, Reykjavik, Iceland, October 2006, *LNCS 4193*, pp. 978–987. Springer.

[50] 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. ACM. To appear (nominated for best paper award).

[51] B. Doerr and C. Klein. Unbiased rounding of rational matrices. In S. Arun-Kumar and N. Garg, eds., *FSTTCS 2006: Foundations of Software Technology and Theoretical Computer Science: 26th International Conference*, Kolkata, India, 2006, *LNCS 4337*, pp. 200–211. Springer.

[52] 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. IEEE. To appear.

[53] B. Doerr, J. Lengler, and D. Steurer. The interval liar game. In T. Asano, ed., *Algorithms and Computation, 17th International Symposium, ISAAC 2006*, Kolkata, India, 2006, *LNCS 4288*, pp. 318–327. Springer.

[54] B. Doerr, F. Neumann, D. Sudholt, and C. Witt. On the influence of pheromone updates in aco algorithms. In *Genetic and Evolutionary Computation Conference (GECCO-2007)*, London, UK, 2007. ACM. To appear (nominated for best paper award).

[55] 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.

[56] A. Eigenwillig, L. Kettner, and N. Wolpert. Snap rounding of Bézier curves. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry (SCG 2007)*, Gyeongju, South Korea, 2007. ACM. To appear.

[57] 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, July 2006, pp. 71–78. ACM.

[58] F. Eisenbrand, S. Funke, A. Karrenbauer, and D. Matijevic. Energy-aware stage illumination. In *Proceedings of the 21st Annual Symposium on Computational Geometry, (SCG05)*, Pisa, Italy, 2005, pp. 336–346. ACM.

[59] F. Eisenbrand, S. Funke, A. Karrenbauer, J. Reichel, and E. Schömer. Packing a trunk - now with a twist! In S. N. Spencer, ed., *Proceedings SPM 2005 ACM Symposium on Solid and Physical Modeling*, Cambridge, USA, June 2005, pp. 197–206. ACM.

[60] K. Elbassioni. Finding all minimal infrequent multi-dimensional intervals. In J. R. Correa, A. Hevia, and M. A. Kiwi, eds., *LATIN 2006: Theoretical Informatics, 7th Latin American Symposium*, Valdivia, Chile, 2006, vol. 3887, pp. 423–434. Springer.

[61] K. Elbassioni. On the complexity of the multiplication method for monotone cnf/dnf dualization. In Y. Azar and T. Erlebach, eds., *Algorithms - ESA 2006, 14th Annual European Symposium*, Zürich, Switzerland, 2006, *LNCS 4168*, pp. 340–351. Springer.

[62] K. Elbassioni, A. Fishkin, N. H. Mustafa, and R. Sitters. Approximation algorithms for euclidean group tsp. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, eds., *Automata, languages and programming, 32nd International Colloquim, ICALP 2005*, Lisbon, Portugal, 2005, *LNCS 3580*, pp. 1115–1126. Springer.

[63] K. Elbassioni, A. V. Fishkin, and R. Sitters. On approximating the TSP with intersecting neighborhoods. In T. Asano, ed., *Algorithms and Computation, 17th International Symposium, ISAAC 2006*, Kolkata, India, 2006, *LNCS 4288*, pp. 213–222. Springer.

[64] K. Elbassioni, I. Katriel, M. Kutz, and M. Mahajan. Simultaneous matchings. In X. Deng and D. Du, eds., *Algorithms and computation, 16th International Symposium, ISAAC 2005*, Sanya, Hainan, China, 2005, *LNCS 3827*, pp. 106–115. Springer.

[65] K. Elbassioni and N. H. Mustafa. Conflict-free colorings of rectangle ranges. In B. Durand and W. Thomas, eds., *STACS 2006, 23rd Annual Symposium on Theoretical Aspects of Computer Science*, Marseille, France, February 2006, *LNCS 3884*, pp. 254–263. Springer.

[66] A. Elmasry and K. Elbassioni. Output-sensitive algorithms for enumerating and counting simplices containing a given point in th plane. In *17th Canadian Conference on Computational Geometry (CCCG'05)*, Windsor, Canada, 2005, no. 17, pp. 248–251. University of Windsor.

[67] 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 (GECCO-2007)*, London, UK, 2007. ACM. To appear.

[68] T. Friedrich, N. Hebbinghaus, and F. Neumann. Rigorous analyses of simple diversity mechanisms. In D. Thierens, ed., *Genetic and Evolutionary Computation Conference (GECCO-2007)*, London, UK, 2007. ACM. To appear (nominated for best paper award).

[69] S. Funke. Topological hole detection in wireless sensor networks and its applications. In *3rd ACM/SIGMOBILE International Workshop on foundations of Mobile Computing (DIAL-M-POMC)*, Cologne, Germany, 2005, pp. 44–53. ACM.

[70] S. Funke, L. Guibas, A. Nguyen, and Y. Wang. Distance-sensitive information brokerage in sensor networks. In P. B. Gibbons, T. F. Abdelzaher, J. Aspnes, and R. Rao, eds., *Distributed Computing in Sensor Systems, Second IEEE International Conference, DCOSS 2006*, San Francisco, USA, 2006, *LNCS 4026*, pp. 234–251. Springer.

[71] S. Funke, A. Kesselmann, U. Meyer, and M. Segal. A simple improved distributed algorithm for minimum CDS in unit disk graphs. In *1st IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2005)*, Montreal, Canada, 2005, vol. 2, pp. 220–223. IEEE.

[72] S. Funke and C. Klein. Hole detection or: "how much geometry hides in connectivity?". In N. Amenta and O. Cheong, eds., *Proceedings of the 22nd Annual Symposium on Computational Geometry, SCG'06*, Sedona, Arizona, USA, 2006, pp. 377–385. ACM.

[73] S. Funke and S. Laue. Bounded-hop energy-efficient broadcast in low-dimensional metrics via coresets. In W. Thomas and P. Weil, eds., *24th Annual Symposium on Theoretical Aspects of Computer Science (STACS 2007)*, Aachen, Germany, 2007, *LNCS 4393*, pp. 272–283. Springer.

[74] S. Funke, T. Malamatos, D. Matijevic, and N. Wolpert. (approximate) conic nearest neighbors and the induced voronoi diagram. In *18th Canadian Conference on Computational Geometry*, Kingston, Canada, 2006, pp. 23–26. School of Computing, Queen's University.

[75] S. Funke and N. Milosavljevic. Infrastructure-establishment from scratch in wireless sensor networks. In V. K. Prasanna, S. Iyengar, P. Spirakis, and M. Welsh, eds., *Distributed computing in sensor systems, First IEEE International Conference, DCOSS 2005*, Marina Del Rey, USA, 2005, *LNCS 3560*, pp. 354–367. Springer.

[76] S. Funke and N. Milosavljevic. Guaranteed-delivery geographic routing under uncertain node locations. In *11th IEEE Conference on Computer Communication (INFOCOM)*, Anchorage, USA, 2007. IEEE. to appear.

[77] S. Funke and N. Milosavljevic. Network sketching or: "how much geometry hides in connectivity? - part ii". In *Proceedings of the eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-07)*, New Orleans, USA, January 2007, pp. 958–967. SIAM.

[78] A. Grigoriev, J. v. Loon, R. Sitters, and M. Uetz. How to sell a graph: Guidelines for graph retailers. In F. V. Fomin, ed., *32nd International Workshop on Graph-Theoretical Concepts in Computer Science, WG 2006*, Bergen, Norway, October 2006, *LNCS 4271*, pp. 125–136. Springer.

[79] 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, June 2005, pp. 163–174. ACM.

[80] R. Hariharan, K. Telikepalli, and K. Mehlhorn. A faster deterministic algorithm for minimum cycle bases in directed graphs. 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. 250–261. Springer.

[81] K. Kaligosi, K. Mehlhorn, J. I. Munro, and P. Sanders. Towards optimal multiple selection. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, eds., *Automata, languages and programming, 32nd International Colloquim, ICALP 2005*, Lisbon, Portugal, 2005, *LNCS 3580*, pp. 103–114. Springer.

[82] K. Kaligosi and P. Sanders. How branch mispredictions affect quicksort. In Y. Azar and T. Erlebach, eds., *Algorithms - ESA 2006, 14th Annual European Symposium*, Zürich, Switzerland, 2006, *LNCS 4168*, pp. 780–791. Springer.

[83] T. Kavitha, K. Mehlhorn, and D. Michail. New approximation algorithms for minimum cycle bases of graphs. In W. Thomas and P. Weil, eds., *24th Annual Symposium on Theoretical Aspects of Computer Science (STACS 2007)*, Aachen, Germany, February 2007, *LNCS 4393*, pp. 512–523. Springer.

[84] A. Kesselmann and D. Kowalski. Fast distributed algorithm for convergecast in ad hoc geometric radio networks. In *Second Annual Conference on Wireless On-demand Network Systems and Services (WONS'05)*, St. Moritz, Switzerland, 2005, pp. 119–124. IEEE.

[85] L. Kettner. Reference counting in library design—optionally and with union-find optimization. In A. Lumsdaine and S. Schupp, eds., *Library-Centric Software Design (LCSD'05)*, San Diego, CA, USA, October 2005, pp. 1–10. Department of Computer Science, Texas A&M University.

[86] L. Kettner. Reference counting in library design—optionally and with union-find optimization. In D. Musser and J. Siek, eds., *Proceedings of the First International Workshop on Library-Centric Software Design, LCSD'05*, San Diego, CA, USA, 2006, *Technical Report*, vol. 06-12, pp. 34–43. Rensselaer Polytechnic Institute, Computer Science Department.

[87] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, and V. Gurvich. Generating cut conjunctions and bridge avoiding extensions in graphs. In X. Deng and D.-Z. Du, eds., *Algorithms and computation, 16th International Symposium, ISAAC 2005*, Sanya, Hainan, China, December 2005, *LNCS 3827*, pp. 156–165. Springer.

[88] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, and V. Gurvich. Generating all vertices of a polyhedron is hard. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'06*, Miami, FL, USA, January 2006, pp. 758–765. ACM / SIAM.

[89] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, V. Gurvich, and K. Makino. Enumerating spanning and connected subsets in graphs and matroids. In Y. Azar and T. Erlebach, eds., *Algorithms - ESA 2006, 14th Annual European Symposium*, Zürich, Switzerland, 2006, *LNCS 4168*, pp. 444–455. Springer.

[90] L. Khachiyan, E. Boros, K. Elbassioni, and V. Gurvich. Generating all minimal integral solutions to monotone $\wedge, \vee$-systems of linear, transversal and polymatroid inequalities. In J. Jedrzejowicz and A. Szepietowski, eds., *Mathematical Foundations of Computer Science 2005, 30th International Symposium, MFCS 2005*, Gdansk, Poland, 2005, *LNCS 3618*, pp. 556–567. Springer.

[91] L. Khachiyan, E. Boros, K. Elbassioni, and V. Gurvich. A new algorithm for the hypergraph transversal problem. In L. Wang, ed., *Computing and combinatorics, 11th Annual International Conference, COCOON 2005*, Kunming, China, August 2005, *LNCS 3595*, pp. 767–776. Springer.

[92] 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.

[93] R. Klein and M. Kutz. The density of iterated crossing points and a gap result for triangulations of finite point sets. In N. Amenta and O. Cheong, eds., *Proceedings of the 22nd Annual Symposium on Computational Geometry (SCG06)*, Sedona, Arizona, USA, 2007, pp. 264–272. ACM.

[94] A. Kovács. Fast monotone 3-approximation algorithm for scheduling related machines. In G. S. Brodal and S. Leonardi, eds., *Algorithms - ESA 2005: 13th Annual European Symposium*, Mallorca, Spain, 2005, *LNCS 3669*, pp. 616–627. Springer.

[95] A. Kovács. Polynomial time preemptive sum-multicoloring on paths. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, eds., *Automata, languages and programming, 32nd International Colloquim, ICALP 2005*, Lisbon, Portugal, July 2005, *LNCS 3580*, pp. 840–852. Springer.

[96] A. Kovács. Tighter approximation bounds for LPT scheduling in two special cases. In T. Calamoneri, I. Finocchi, and G. F. Italiano, eds., *Algorithms and Complexity, 6th Italian Conference, CIAC 2006*, Rome, Italy, May 2006, *LNCS 3998*, pp. 187–198. Springer.

[97] L. Kowalik. Approximation scheme for lowest outdegree orientation and graph density measures. In T. Asano, ed., *Algorithms and Computation, 17th International Symposium, ISAAC 2006*, Kolkata, India, 2006, *LNCS 4288*, pp. 557–566. Springer.

[98] L. Kowalik. Improved edge coloring with three colors. In F. V. Fomin, ed., *Graph-Theoretic Concepts in Computer Science, 32nd International Workshop, WG 2006*, Bergen, Norway, 2006, *LNCS 4271*, pp. 90–101. Springer.

[99] D. Kowalski, P. M. Musial, and A. Shvartsman. Explicit combinatorial structures for cooperative distributed algorithms. In *25th IEEE International Conference on Distributed Computing Systems (ICDCS'05)*, Ohio, USA, 2005, pp. 49–58. IEEE.

[100] M. Kutz. Weak positional games on hypergraphs of rank three. In S. Felsner, ed., *2005 European Conference on Combinatorics, Graph Theory and Applications (EuroComb '05)*, Berlin, Germany, September 2005, *DMTCS Proceedings*, vol. AE, pp. 31–36. DMTCS. to appear.

[101] M. Kutz. Computing shortest non-trivial cycles on orientable surfaces of bounded genus in almost linear time. In N. Amenta and O. Cheong, eds., *Proceedings of the 22nd Annual Symposium on Computational Geometry, SCG'06*, Sedona, Arizona, USA, 2006, pp. 430–437. ACM.

[102] M. Kutz and A. Pór. Angel, devil, and king. In L. Wang, ed., *Computing and Combinatorics, 11th Annual International Conference, COCOON 2005*, Kunming, China, 2005, *LNCS 3595*, pp. 925–934. Springer.

[103] M. Kutz and P. Schweitzer. ScrewBox: a randomized certifying graph non-isomorphism algorithm. In D. Applegate and G. Brodal, eds., *9th Workshop on Algorithm Enginneering and Experiments (ALENEX'07)*, New Orleans, USA, 2007. SIAM.

[104] Z. Lotker, D. Majumdar, N. Narayanaswamy, and I. Weber. Sequences characterizing k-trees. In D. Z. Chen and D. T. Lee, eds., *Computing and Combinatorics, 12th Annual International Conference, COCOON 2006*, Taipei, Taiwan, 2006, *LNCS 4112*, pp. 216–225. Springer.

[105] T. Malamatos. Lower bounds for expected-case planar point location. In *17th Canadian Conference on Computational Geometry (CCCG'05)*, Windsor, Canada, 2005, pp. 191–194. University of Windsor.

[106] J. Maue, P. Sanders, and D. Matijevic. Goal directed shortest path queries using precomputed cluster distances. In C. Alvarez and M. J. Serna, eds., *Experimental Algorithms, 5th International Workshop, WEA 2006*, Cala Galdana, Menorca, Spain, 2006, *LNCS 4007*, pp. 316–327. Springer.

[107] K. Mehlhorn and D. Michail. Implementing minimum cycle basis algorithms. In S. Nikoletseas, ed., *Experimental and Efficient Algorithms, 4th InternationalWorkshop, WEA 2005*, Santorini, Greece, May 2005, *LNCS 3503*, pp. 32–43. Springer.

[108] 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.

[109] U. Meyer and S. Wetzel. Introducing history-enriched security context transfer to enhance the security of subsequent handover. In *4th IEEE Conference on Pervasive Computing and Communications Workshops (PerCom 2006 Workshops)*, Pisa, Italy, 2006, pp. 277–282. IEEE.

[110] U. Meyer and N. Zeh. I/O-efficient undirected shortest paths with unbounded edge lengths. In Y. Azar and T. Erlebach, eds., *Algorithms - ESA 2006, 14th Annual European Symposium*, Zurich, Switzerland, September 2006, *LNCS 4168*, pp. 540–551. Springer.

[111] 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 2007)*, Gyeongju, South Korea, 2007. ACM. To appear.

[112] C. W. Mortensen and S. Pettie. The complexity of implicit and space-efficient priority queues. In F. Dehne, A. López-Ortiz, and J.-R. Sack, eds., *Algorithms and data structures, 9th International Workshop, WADS 2005*, Waterloo, Canada, 2005, *LNCS 3608*, pp. 49–60. Springer.

[113] N. H. Mustafa and S. Ray. An optimal generalization of the centerpoint theorem and its extensions. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry (SCG 2007)*, Gyeongju, South Korea, 2007. ACM.

[114] N. H. Mustafa and S. Ray. 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 2007)*, Gyeongju, South Korea, 2007. ACM.

[115] K. Paluch. A new approximation algorithm for multidimensional rectangle tiling. In T. Asano, ed., *Algorithms and Computation, 17th International Symposium, ISAAC 2006*, Kolkata, India, 2006, pp. 712–721. Springer.

[116] S. Pettie. Sensitivity analysis of minimum spanning trees in sub-inverse-Ackermann time. In X. Deng and D. Du, eds., *Algorithms and computation, 16th International Symposium, ISAAC 2005*, Sanya, Hainan, China, 2005, *LNCS 3827*, pp. 964–973. Springer.

[117] 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.

[118] S. Schmitt. The diamond operator - implementation of exact real algebraic numbers. 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. 355–366. Springer.

[119] R. Seidel and N. Wolpert. On the exact computation of the topology of real algebraic curves. In J. S. B. Mitchell and G. Rote, eds., *Proceedings of the 21st ACM Symposium on Computational Geometry*, Pisa, Italy, 2005, pp. 107–115. ACM.

[120] G. Weikum, H. Bast, G. Canright, D. Hales, C. Schindelhauer, and P. Triantafillou. Towards self-organizing query routing and processing for peer-to-peer web search. In C. Schindelhauer, ed., *Workshop on Peer-to-peer Data Management in the Complex Systems Perspective*, Paris, France, November 2005, pp. 7–24. University of Paderborn, Heinz Nixdorf Institute.

[121] 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, November 2007.

Theses

[1] D. Ajwani. Design, implementation and experimental study of external memory BFS algorithms. Masters thesis, Universität des Saarlandes, 2005.

[2] A. Baus. Symbolic constraints in linear integer programming. Masters thesis, Universität des Saarlandes, March 2005.

[3] M. Caroli. Exakte Arrangement-Berechnung gedrehter Quadratischer Kurven. Bachelor thesis, Universität des Saarlandes, April 2006.

[4] M. Caroli. Evaluation of a generic method for analyzing controlled-perturbation algorithms. Masters thesis, Universität des Saarlandes, March 2007.

[5] R. Dementiev. *Algorithm Engineering for Large Data Sets*. Phd thesis, Universität des Saarlandes, December 2006.

[6] B. Doerr. *Integral Approximation*. Habilitation thesis, Christian-Albrechts-Universität zu Kiel, 2005.

[7] F. Ebert. Benchmark data sets for conic arrangements. Bachelor thesis, Universität des Saarlandes, November 2005.

[8] P. Emeliyanenko. Visualization of points and segments of real algebraic plane curves. Masters thesis, Universität des Saarlandes, February 2007.

[9] M. Fouz. Hereditary discrepancy in different numbers of colors. Bachelor thesis, Universität des Saarlandes, October 2006.

[10] T. Friedrich. Deterministic random walks on infinite grids. Masters thesis, Friedrich-Schiller-Universität Jena, December 2005. Mathematik-Diplomarbeit, eingereicht an der Friedrich-Schiller-Universität Jena.

[11] P. Hachenberger. *Boolean Operations on 3D Selevctive Nef Complexes: Data Structure, Algorithms Optimized Implementation, Experiments and Applications.* Phd thesis, Universität des Saarlandes, December 2006.

[12] D. Johannsen. Sampling rooted 3-connected planar graphs in deterministic polynomial time. Masters thesis, Humboldt-Universität zu Berlin, April 2006.

[13] M. Kamran Azam. Branch-and-cut techniques for generalized asymmetric traveling salesman problem. Masters thesis, Universität des Saarlandes, September 2005.

[14] M. Kerber. Analysis of real algebraic plane curves. Masters thesis, Universität des Saarlandes, September 2006.

[15] C. Lennerz. *Distance Computation for Extended Quadratic Complexes.* Phd thesis, Universität des Saarlandes, January 2005.

[16] J. Maue. A goal-directed shortest path algorithm using precomputed cluster distances. Masters thesis, Universität des Saarlandes, June 2006.

[17] D. Michail. *Minimum Cycle Basis, Algorithms and Applications.* Phd thesis, Universität des Saarlandes, July 2006.

[18] R. Newo Kenmogne. Understanding lsi via the truncated term-term matrix. Masters thesis, Universität des Saarlandes, May 2005.

[19] S. Pohl. Exact integer linear programming with bounded variables in a branch- and cut algorithm. Masters thesis, Universität des Saarlandes, March 2006.

[20] E. Pyrga. Shortest paths in time-dependent networks and their applications. Masters thesis, Universität des Saarlandes, January 2005.

[21] I. Rauf. Earliest arrival flows with multiple sources. Masters thesis, Universität des Saarlandes, March 2005.

[22] J. Reichel. *Combinatorial Approaches to Trunk Packing.* Phd thesis, Universität des Saarlandes, July 2006.

[23] D. Schmitt. Implementierung einer Überlagerung von konvexen Arrangements der Kugeloberfläche. Bachelor thesis, Universität des Saarlandes, August 2006.

[24] D. Schultes. Fast and exact shortest path queries using highway hierachies. Masters thesis, Universität des Saarlandes, August 2005.

[25] D. Steurer. An asymptomic approximation scheme for multigraph edge coloring. Masters thesis, Universität des Saarlandes, August 2006.

[26] D. Steurer. Tight bounds on the min-max boundary decomposition cost of weighted graphs. Bachelor thesis, Universität des Saarlandes, May 2006.

[27] D. Weber. Solving large sparse linear systems exactly. Masters thesis, Universität des Saarlandes, August 2006.

[28] C. Weinand. Fill-in reduction while solving large, sparse linear systems with graph theoretical methods. Masters thesis, Universität des Saarlandes, June 2005.

Technical reports

[1] H. Bast, D. Majumdar, R. Schenkel, M. Theobald, and G. Weikum. IO-Top-k: Index-access optimized top-k query processing. Research Report MPI-I-2006-5-002, Max-Planck-Institut for Informatics, Saarbrücken, Germany, March 2006.

[2] H. Bast, I. Weber, and C. W. Mortensen. Output-sensitive autocompletion search. Research Report MPI-I-2006-1-007, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, June 2006.

[3] S. Baswana and K. Telikepalli. Improved algorithms for all-pairs approximate shortest paths in weighted graphs. Research Report MPI-I-2005-1-003, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, September 2005.

[4] E. Berberich and M. Hemmer. Prototype implementation of the algebraic kernel. Technical Report ACS-TR-121202-01, University of Groningen, Groningen, Netherlands, April 2006.

[5] E. Berberich, M. Hemmer, M. Karavelas, S. Pion, M. Teillaud, and E. Tsigaridas. Interface specification of algebraic kernel. Technical Report ACS-TR-123101-01, University of Groningen, Groningen, The Netherlands, April 2006.

[6] J. Cooper, B. Doerr, J. Spencer, and G. Tardos. Deterministic random walks on the integers, 2006.

[7] R. Dementiev, L. Kettner, and P. Sanders. STXXL: Standard template library for XXL data sets. Technical Report 2005/18, Fakultät für Informatik, University of Karlsruhe, Karlsruhe, Germany, 2005.

[8] B. Doerr and M. Fouz. Hereditary discrepancies in different numbers of colors ii, November 2006.

[9] B. Doerr and M. Gnewuch. Construction of low-discrepancy point sets of small size by bracketing covers and dependent randomized rounding. Research Report 06-14, University Kiel, Kiel, 2006.

[10] A. Eigenwillig, L. Kettner, and N. Wolpert. Snap rounding of Bézier curves. Research Report MPI-I-2006-1-005, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, December 2006.

[11] A. Eigenwillig, L. Kettner, and N. Wolpert. Snap rounding of Bézier curves. Technical Report ACS-TR-121108-01, Max-Planck-Institut für Informatik, 66123 Saarbrücken, Germany, March 2007.

[12] K. M. Elbassioni. On the complexity of monotone boolean duality testing. Technical Report DIMACS TR: 2006-01, DIMACS, Piscataway, NJ ,USA, 2006.

[13] S. Funke, C. Klein, K. Mehlhorn, and S. Schmitt. Controlled perturbation for delaunay triangulations. Technical Report ACS-TR-121103-03, Algorithms for Complex Shapes with certified topology and numerics, Instituut voor Wiskunde en Informatica, Groningen, NETHERLANDS, 2006.

[14] S. Funke, S. Laue, R. Naujoks, and L. Zvi. Power assignment problems in wireless communication. Research Report MPI-I-2006-1-004, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, December 2007.

[15] N. Hebbinghaus. Discrepancy of sums of two arithmetic progressions. ArXiv math.NT/0703108, Cornell University (ArXiv), ArXiv, March 2007.

[16] I. Katriel, M. Kutz, and M. Skutella. Reachability substitutes for planar digraphs. Research Report MPI-I-2005-1-002, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, March 2005.

[17] M. Kerber. Division-free computation of subresultants using bezout matrices. Research Report MPI-I-2006-1-006, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, May 2006.

[18] F. Neumann and C. Witt. Ant colony optimization and the minimum spanning tree problem. Technical Report TR06-143, Electronic Colloquium on Computational Complexity, http://eccc.hpi-web.de/eccc/, November 2006.

13 The Programming Logics Group (D2)

13.1 Personnel

Director
Prof. Dr. Harald Ganzinger († 2004)

Acting Director
Prof. Dr. Thomas Lengauer, PhD

Senior Research Scientist
Prof. Dr. Andreas Podelski (–June 2006)

Researchers
Dr. Peter Baumgartner (–November 2005)
Dr. Hans de Nivelle (–February 2007)
Dr. Jörn Freiheit
Dr. Jörg Hoffmann (–August 2006)
Dr. Chin Soon Lee (–December 2005)
Dr. Patrick Maier (–September 2006)
Dr. Virgile Prevesto (–June 2006)
Dr. Stefan Ratschan (–September 2006)
Dr. Zhikun She (–December 2006)
Dr. Viorica Sofronie-Stokkermans[1]
Dr. Uwe Waldmann[1]

PhD Students
Thomas Hillenbrand[1]
Carsten Ihlemann[1]
Swen Jacobs[1]
Yevgeny Kazakov (–October 2005)
Alexander Malkis (–October 2006)
Stefan Maus (–June 2006)
Ruzika Piskac (–September 2006)
Andrey Rybalchenko (–December 2006)
Ina Schäfer (–October 2006)

[1]Thomas Hillenbrand, Carsten Ihlemann, Swen Jacobs, Viorica Sofronie-Stokkermans, and Uwe Waldmann moved to Research Group 1 "Automation of Logic" in September 2005. With the exception of Subsection 13.3.2, their work is reported in Chapter 17.

Nassim Seghir (–September 2006)
Silke Wagner
Thomas Wies (–September 2006)

Secretaries
Brigitta Hansen
Veronika Weinand

13.2 Visitors

In the time period from January 2006 to December 2006, the following researchers visited our group:

Jerzy Marcinkowski	11.02.05–16.02.05	University of Wrocław
Oded Maler	12.02.05–21.02.05	Verimag, Gieres
Carmel Domshlak	10.07.05–17.07.05	Technion – Isreal Institute of Technology
Andrei Voronkov	19.09.05	University of Manchester
Oded Maler	22.10.05–03.11.05	Verimag, Gieres
Dino Distefano	05.12.05–07.12.05	University of London
Peter Revesz	01.12.05–31.10.06	University of Nebraska-Lincoln
Bruno Berstel	18.12.05–21.12.05	ILOG SA, Gentilly
Michel Leconte	19.12.05–21.12.05	ILOG SA, Gentilly
Anai Hirokazu	24.02.06–28.02.06	Fujitsu Labaratories
Henning Burchardt	19.02.06–22.02.06	University of Oldenburg
Tino Teige	28.02.06–03.03.06	University of Oldenburg
Werner Damm	28.02.06–03.03.06	University of Oldenburg
Ernst-Rüdiger Olderog	20.03.06–24.03.06	University of Oldenburg
Goran Frehse	14.12.06–15.12.06	Verimag, Gieres

13.3 Instantiation-Based Theorem Proving

Instantiation-based theorem proving refers to a family of methods for first-order logic theorem proving. They share the principle of carrying out proof search by maintaining a set of instances of input formulas, typically clauses, and analyzing it for propositional satisfiability until completion. While these two aspects “instantiation” and “propositional satisfiability” are underlying most methods for automated theorem proving, instantiation-based calculi separate them. This allows to explore new ways of their interleaving. The spectrum ranges from using Sat procedures as black boxes, thereby benefiting from recent remarkable improvements in the performance of propositional Sat procedures, to fine-grained methods of interleaving a la resolution [4], a la Tableaux [10] or a la Davis/Putnam [2].

13.3.1 Model Evolution

Investigator: Peter Baumgartner

The Model Evolution calculus [2], one of our own methods, is a procedure that stays close to the propositional DPLL procedure and lifts it to first-order logic. One of the main motivations for this approach was the possibility of migrating to the first-order level some of those very effective search techniques developed by the SAT community for the DPLL procedure.

We have continued our work on the Model Evolution calculus. We have improved our implementation, the Darwin system [1], and we have advanced the theory by integrating inference rules for equality [3] by adapting the Bachmair/Ganzinger framework for the superposition calculus.

13.3.2 Comparison of Instantiation-Based Methods

Investigators: Swen Jacobs and Uwe Waldmann

We started investigations into the relations between different instantiation-based methods known from the literature. Such investigations are important, as they not only help to clarify conceptual differences and similarities, but may also provide exact technical results as a basis to identify qualitative differences and commonalities.

Surprisingly, almost no such exact technical results have been published. Indeed, when starting such comparisons at a detailed, technical level it soon becomes clear that many of the commonalities intuitively expected or informally observed do not hold, or can be established on a technical level only with difficulties. This is one of the conclusions in [8], which is the first attempt of its kind to rigorously compare instantiation-based methods. More specifically, [8] covers the Disconnection Tableau Calculus [10], Primal Partial Instantiation [7], and Resolution-Based Instance Generation [4]. We investigate the relationship between these calculi and answer the question to what extent refutation or consistency proofs in one calculus can be simulated in another one.

In [9], we extend our results on this comparison. For two calculi which are equivalent up to their fairness criteria (Primal Partial Instantiation and a version of Resolution-Based Instance Generation), we give a fairness criterion that preserves completeness and generalizes those that come with the respective calculi.

13.3.3 Theory Reasoning in Instantiation-Based Calculi

Investigators: Harald Ganzinger and Konstantin Korovin

The idea of instantiation-based reasoning is to reduce undecidable first-order problems to decidable instances of a given problem in a complete way. In the basic methods, first-order problems are reduced to propositional logic by instantiating them to ground formulae. The approach has been extended for several calculi to handle the theory of equality with special proof rules (e.g., [3, 5]. The explicit integration of a given theory allows for much more efficient handling of problems in that theory than the general first-order reasoning of the calculus itself.

A natural extension of this approach is to allow the reduction to arbitrary theories for which complete reasoners exist. In [6], we developed an integration of theory reasoners into

the instantiation framework. The integration is done in a black-box style, such that different theories can be integrated in a uniform way. With this approach, existing satisfiability solvers for decidable theories can be used in the general first-order reasoning framework of instantiation-based reasoning.

The combination of general first-order reasoning with reasoners for specific theories allows for efficient solving of problems in these theories without losing the expressiveness of first-order logic.

References

- [1] P. Baumgartner, A. Fuchs, and C. Tinelli. Implementing the model evolution calculus. *International Journal on Artificial Intelligence Tools*, 15(1):21–52, 2006.

[2] P. Baumgartner and C. Tinelli. The model evolution calculus. In F. Baader, ed., *CADE-19 – The 19th International Conference on Automated Deduction*, 2003, LNAI 2741, pp. 350–364. Springer.

- [3] P. Baumgartner and C. Tinelli. The model evolution calculus with equality. In R. Nieuwenhuis, ed., *Automated deduction - CADE-20, 20th International Conference on Automated Deduction*, Tallinn, Estonia, 2005, *LNAI 3632*, pp. 392–408. Springer.

[4] 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.

[5] H. Ganzinger and K. Korovin. Integration of equational reasoning into instantiation-based theorem proving. In J. Marcinkowski and A. Tarlecki, eds., *Computer science logic, 18th International Workshop CSL 2004, 13th Annual Conference of the EACSL*, Karpacz, Poland, September 2004, *LNCS 3210*, pp. 71–84. Springer.

- [6] 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.

[7] J. N. Hooker, G. Rago, V. Chandru, and A. Rivastava. Partial instantiation methods for inference in first-order logic. *Journal of Automated Reasoning*, 28:371–396, 2002.

- [8] S. Jacobs and U. Waldmann. Comparing instance generation methods for automated reasoning. In B. Beckert, ed., *Automated Reasoning with Analytic Tableaux and Related Methods, International Conference, TABLEAUX 2005*, Koblenz, Germany, 2005, *LNCS 3702*, pp. 153–168. Springer.

- [9] S. Jacobs and U. Waldmann. Comparing instance generation methods for automated reasoning. *Journal of Automated Reasoning*, 38:57–78, April 2007.

[10] R. Letz and G. Stenz. Proof and model generation with disconnection tableaux. In R. Nieuwenhuis and A. Voronkov, eds., *Logic for Programming, Artificial Intelligence, and Reasoning: 8th International Conference*, 2001, LNAI 2250, pp. 142–156. Springer.

13.4 Software Verification

Computer systems (e.g. operating systems, web servers, mail servers, database engines, etc) are usually constructed from a set of components that we expect will always terminate and

deliver requested results. Cases where these components unexpectedly do not response to requests can lead to the loss of system functionality, or even make it complete useless for any tasks.

We develop algorithms that allow one to automatically identify and analyze aspects of a given software system w.r.t. the fulfillment of functionality that is necessary for failure-free operation. Such analysis is an extremely difficult task, as modern software is one of the most complicated constructions built by programmers/engineers. We tackle the inherent analysis complexity by developing algorithms that employ logical reasoning about programs, their computations, and functionality realized by these computations.

The logical reasoning about software allows us to systematically find the right level of abstraction automatically, when analyzing a program. The abstraction process is required to achieve necessary applicability. During this process, it is crucial to keep enough information about the underlying program in order to capture the desired properties. Moreover, it is indispensable to abstract away nonessential details and specifics, to prevent them from hindering the efficiency of the analysis. The main challenge is to find the right level of abstraction automatically.

13.4.1 Path Invariants

Investigators: Andrey Rybalchenko in collaboration with Dirk Beyer, Thomas Henzinger, and Rupak Majumdar

The success of software verification depends on the ability to find a suitable abstraction of a program automatically. We propose a new method for automated abstraction refinement, which overcomes the inherent limitations of predicate discovery schemes. In such schemes, the cause of a false positive is identified as an infeasible error path, and the abstraction is refined in order to remove that path. By contrast, we view the cause of a false positive – the "spurious counterexample" – as a full-fledged program, whose control-flow graph may contain loops of the original program and represent unbounded computations [8]. The advantages of using such *path programs* as counterexamples for abstraction refinement are twofold. First, we can bring the whole machinery of program analysis to bear on path programs: specifically, we use abstract interpretation in the form of constrained-based invariant generation [7] to automatically infer invariants of path programs – so-called *path invariants*. Second, we use path invariants for abstraction refinement in order to remove not one infeasibility at a time, but to remove at once all infeasible error computations that are represented by a path program. Unlike predicate discovery schemes, our method handles loops without unrolling them; it infers abstractions that involve universal quantification and naturally incorporates disjunctive invariants.

13.4.2 Proving Liveness

Investigators: Andreas Podelski and Andrey Rybalchenko in collaboration with Byron Cook

Discovery of practical algorithms for proving that software systems eventually deliver requested results had been an open problem since 1970s. One of the major obstacles was the lack of adequate abstraction techniques.

To provide a solution to this problem, we developed a new kind of logical assertions, called transition invariants [4]. They overcome the inherent limitations of traditional methods for proving liveness (which is the technical term for the eventual delivery of results). We provided an abstraction refinement algorithm that determines an adequate level of abstraction automatically [1]. We also developed necessary ingredients for designing an algorithm for proving liveness of large code fragments written in C programming language.

Our theoretical investigations [4, 1] provided foundations for the development of practical verification tools for liveness properties. We implemented industrial-scale tools [3] that are successfully applied for the verifications of critical fragments of systems code, in cooperation with Microsoft Research laboratory and within the Verisoft project. The major applications include automated liveness proofs for dispatch routines in device drivers, and for low level routines in operating system kernel [2, 9]. The liveness proofs for these components guarantee that the operating system will never stop providing its services for the user.

The application area of systems code motivates the development of liveness checking methods for concurrent programs, since many system components are executed in a concurrent setting. We developed a method for proving termination of threads in multi-threaded programs, which achieves its practical applicability by applying counterexample-guided refinement techniques to infer environment assumptions for the thread under consideration [10].

13.4.3 ARMC: Abstraction Refinement Model Checker

Investigators: Andreas Podelski and Andrey Rybalchenko

Software model checking with abstraction refinement is emerging as a practical approach to verify industrial software systems. Its distinguishing characteristics lie in the way it applies logical reasoning to deal with abstraction. It is therefore natural to investigate whether and how the use of a constraint-based programming language may lead to an elegant and concise implementation of a practical tool [6]. Using a Prolog system together with Constraint Logic Programming extensions as the implementation platform of our choice we have built such a tool, called ARMC (for Abstraction Refinement Model Checking), which has already been used for practical verification [5].

References

- [1] B. Cook, A. Podelski, and A. Rybalchenko. Abstraction-refinement for termination. In C. Hankin and I. Siveroni, eds., *Static analysis, 12th International Symposium, SAS 2005*, London, UK, September 2005, *LNCS 3672*, pp. 87–101. Springer.
- [2] B. Cook, A. Podelski, and A. Rybalchenko. Termination proofs for systems code. In *PLDI 2006, Proceedings of the ACM SIGPLAN 2006 Conference on Programming Language Design and Implementation*, Ottawa, Ontario, Canada, 2006, *ACM SIGPLAN Notices*, vol. 41, pp. 415–426. ACM.
- [3] B. Cook, A. Podelski, and A. Rybalchenko. Terminator: Beyond safety. In T. Ball and R. B. Jones, eds., *Computer aided verification, 18th International Conference, CAV 2006*, Seattle, WA, USA, 2006, *LNCS 4144*, pp. 415–418. Springer.

[4] A. Podelski and A. Rybalchenko. Transition invariants. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science, LICS 2004*, Turku, Finland, July 2004, pp. 32–41. IEEE.

- [5] A. Rybalchenko. Model checking duration calculus: a practical approach. In *Theoretical Aspects of Computing - ICTAC 2006, Third International Colloquium*, Tunis, Tunisia, 2006, *LNCS 4281*, pp. 332–346. Springer.
- [6] A. Rybalchenko. Armc: the logical choice for software model checking with abstraction refinement. In *9th International Symposium on Practical Aspects of Declarative Languages (PADL 2007)*, Nice, France, 2007, *LNCS 4281*, pp. 245–259. Springer.
- [7] A. Rybalchenko. Invariant synthesis for combined theories. In A. Podelski, ed., *8th International Conference on Verification, Model Checking, and Abstract Interpretation (VMCAI 2007)*, Nice, France, 2007, *LNCS 4349*, pp. 0–0. Springer.
- [8] A. Rybalchenko. Path invariants. In *ACM SIGPLAN 2007 Conference on Programming Language Design and Implementation*, San Diego, USA, 2007, pp. 0–0. ACM.
- [9] A. Rybalchenko. Proving that programs eventually do something good. In *34th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL 2007)*, Nice, France, 2007, pp. 265–276. ACM.
- [10] A. Rybalchenko. Proving thread termination. In *ACM SIGPLAN 2007 Conference on Programming Language Design and Implementation*, San Diego, USA, 2007, pp. 0–0. ACM.

13.4.4 Geometric Resolution

Investigator: Hans de Nivelle

Geometric Resolution is a new proof search method for first-order logic. Like standard resolution or superposition, geometric resolution is complete for first-order logic with equality. The reasons that we introduced geometric resolution, are the following: long-term goals in mind: **(1)**. We want to obtain a procedure which is better than resolution in finding finite models. Although resolution can be turned into a decision procedure for many subsets of first-order logic, there seems to exist no practical way of ensuring that it terminates in all cases where a finite model exists. For geometric resolution such a strategy exists. Even when resolution terminates, there is no known general method for obtaining a counter model from the final search state. Geometric resolution always constructs a counter model when it terminates without proof. **(2)**. We want to obtain a proof search method that is better in dealing with partial and ill-defined functions than resolution. Most functions that occur in the real world (programming languages for us) are not functions in the mathematical sense. They are typically partial and not fully specified. (Think of the function that returns a minimal element in a priority queue. The queue may be empty, in that case there is nothing to return, or the minimal object not be unique, in which case the implementation is free to choose one) Because in geometric resolution functions are replaced by relations, such ill-defined functions can be dealt with in a natural fashion. **(3)**. Obtain more efficient proof search procedures by avoiding the symmetries that are caused by multiple symbolic representations of the same object. (We now become a bit speculative) In first-order logic with equality, the same object usually has many representations. For example, under the equality $s(0) \approx 0$, the object 0 has infinitely many representations of form $s^i(0)$. When $+$ is an associative-commutative operator, the expression $1 + (2 + 3) + 4$ has many representations. In superposition, the use of *redundancy elimination* (in particular rewriting) is able to clean up some of the multiple representations, but not all. No satisfactory method has been

found in the presence of associative-commutative symbols. Because of this, it might be much better to use a proof search method based on interpretations, which does not introduce such multiple representations in the first place.

Geometric resolution uses the following formula format:

$$\forall \overline{x}\ A_1(\overline{x}) \wedge \cdots \wedge A_p(\overline{x}) \wedge x_1 \not\approx x'_1 \wedge \cdots \wedge x_q \not\approx x'_q \rightarrow Z(\overline{x}),$$

in which $p \geq 0,\ q \geq 0$, and the $x_1, x'_1, \ldots, x_q, x'_q \in \overline{x}$ are variables. There are no functions and no constants in the formulas.

The right hand side $Z(\overline{x})$ must have one of the following three forms:

1. The false constant $\bot$.
2. A non-empty disjunction of atoms $B_1(\overline{x}) \vee \cdots \vee B_r(\overline{x})$ with $r > 0$.
3. An existential formula of form $\exists y\ B(\overline{x}, y)$ with $y \in \mathcal{V}$ but $y \notin \overline{x}$. The variable y must occur in $B(\overline{x}, y)$.

Rules of the first type are called *lemmas*. The calculus consists of a combination of model enumeration and lemma generation. Model search starts with the empty model. After that it searches as follows: In case the current structure I is not a model, this must be due to a rule which has the left hand side true, and the right hand side false. In case the rule is a lemma, the model search attempt has failed, and the system backtracks. In case the rule is of the second or third type, the model can be extended, possibly in different ways, and the algorithm has to backtrack through the possible extensions. When the model search algorithm has attempted all possibilities, and none of the branches could be extended to a model, it is possible to construct a new lemma which immediately refutes the model search attempt before the backtracking point.

13.4.5 Verification of a Result Checker for Priority Queues

Investigators: Hans de Nivelle and Ruzica Piskac

In [2] we verified with the assistance of a resolution theorem prover, a datastructure that checks a priority queue. The checking datastructure originates from LEDA [1]. For the verification, we used the theorem prover SATURATE.

In order to do this, we developed a new framework, in which it can be specified up to which level a datastructure meets a certain specification. For example, it is possible to specify that an implementation completely meets a specification, or that an implementation has behaved in accordance with the specification for a certain amount of time. The latter condition was used for the verification of the result checker.

We formally proved that, whenever the checking datastructure checks some object P, and there is a moment t on which P does not behave in accordance with the specification of a priority queue, then there exists a moment $t' > t$ on which the checking datastructure will generate an error.

Ideally, one would like to see $t = t'$, which means that any misbehavior of P is detected immediately. Such a checker is called *online result checker*. However, it has been proven in [1] that every online checker for priority queues would have a computational cost which is

too high in comparison with the priority queue P itself. In contract, the datastructure in [1] has a computational cost which is negligible in comparison to the cost of the priority queue. But as a consequence, it has to be an *offline checker*. This is a result checker that detects incorrect behavior eventually, but not immediately.

The complete verification tree can be found on http://www.mpi-inf.mpg.de/ rpiskac/. It consists of 64 calls to SATURATE.

References

[1] U. Finkler and K. Mehlhorn. Checking priority queues. In *Proceedings of the 10th annual ACM-SIAM symposium on Discrete algorithms (SODA'99)*, 1999, pp. 901–902. Society for Industrial and Applied Mathematics.

- [2] H. de Nivelle and R. Piskac. Verification of an off-line checker for priority queues. In B. K. Aichernig and B. Beckert, eds., *Third IEEE International Conference on Software Engineering and Formal Methods (SEFM 2005)*, Koblenz, 2005, pp. 210–219. IEEE.

13.5 Verification of Hybrid Systems

Investigators: Stefan Ratschan and Zhikun She

We designed a new method for the verification of hybrid systems [3, 5]. The method is based on abstraction refinement where the corresponding abstraction is computed using interval constraint propagation techniques. The advantages of the method are correctness (in contrast to several other methods, our method's correctness is not hampered by rounding errors), generality (the method can handle non-linear differential equations and inequalities, non-linear switching conditions, and even differential-algebraic equalities), and efficiency (the method performs well on a large set of benchmark problems). We introduced several improvements of the basic method [6, 4] and implemented it in the open source software package HSolver (`http://hsolver.sourceforge.net`) based on our constraint solving package RSolver (`http://rsolver.sourceforge.net`) [2].

In addition we studied an undecidable verification problem for discrete-time hybrid systems and proved that a verification technique based on interval arithmetic will terminate with a successful verification, for all robust input problems [1]. A hybrid system is robust wrt. to a property, if the property holds on the system and on all sufficiently small perturbations of the system. Such robustness is one of the basic design requirements in practical applications. Hence our method terminates on all well-designed practical problem instances.

References

- [1] W. Damm, G. Pinto, and S. Ratschan. Guaranteed termination in the verification of ltl properties of non-linear robust discrete time hybrid systems. In D. A. Peled and Y.-K. Tsay, eds., *Automated technology for verification and analysis, Third International Symposium, ATVA 2005*, Taipei, Taiwan, 2005, *LNCS 3707*, pp. 99–113. Springer.

- [2] S. Ratschan. Efficient solving of quantified inequality constraints over the real numbers. *ACM Transactions on Computational Logic*, 7(4):723–748, 2006.

- [3] S. Ratschan and Z. She. Safety verification of hybrid systems by constraint propagation based abstraction refinement. In M. Morari and L. Thiele, eds., *Hybrid Systems: Computation and Control: 8th International Workshop, HSCC 2005*, Zürich, Schweiz, 2005, *LNCS 3414*, pp. 573–589. Springer.

- [4] S. Ratschan and Z. She. Constraints for continuous reachability in the verification of hybrid systems. In J. Calmet, T. Ida, and D. Wang, eds., *Artificial Intelligence and Symbolic Computation, 8th International Conference, AISC 2006*, Beijing, China, September 2006, *LNAI 4120*, pp. 196–210. Springer. This work was partly supported by the German Research Council (DFG) as part of the Transregional Collaborative Research Center "Automatic Verification and Analysis of Complex Systems" (SFB/TR 14 AVACS). See www.avacs.org for more information.

- [5] S. Ratschan and Z. She. Safety verification of hybrid systems by constraint propagation based abstraction refinement. *ACM Transactions in Embedded Computing Systems (TECS)*, 6(1):8, 2007.

- [6] S. Ratschan and J.-G. Smaus. Verification-integrated falsification of non-deterministic hybrid systems. In *2nd IFAC Conference on Analysis and Design of Hybrid Systems*, Alghero, Sardinia, November 2006. Elsevier.

13.5.1 A Model Checker for Region Stability of Hybrid Systems

Investigators: Andreas Podelski and Silke Wagner

Recent technological innovations have caused a considerable interest in the study of dynamical processes having a heterogeneous continuous and discrete nature. Such hybrid systems are characterized by the interaction of continuous parts, governed by differential equations, and by discrete parts, described by finite state machines. Hybrid systems switch between many operating modes where each mode is governed by its own characteristic dynamical laws. Application areas of hybrid systems include automotive, manufacturing, communication networks, aerospace, robotics, traffic control, and chemical processes.

For a large class of correctness properties of hybrid systems, push-button verification methods such as model checking have reached a certain degree of practicality. Those properties have in common that they can be reduced to non-reachability. In contrast, for the class of *stability* properties practical pushdown verification methods have so far been out of reach. These properties can not be reduced to non-reachability. Existing verification methods for stability are based on Lyapunov theory. They all share the drawback that state abstraction techniques are intrinsically not applicable. Since such abstraction techniques (used for over-approximation of state spaces and based formally on abstract interpretation) have been crucial to obtain scalability in existing model checkers for hybrid systems, it seems useful to investigate new verification methods for stability.

We have developed a model checking method and tool that integrates state abstraction techniques for the automatic proof of a particular stability property called *region stability* [1, 2]. Region stability means that for each trajectory of the hybrid system there exists a point of time after which it never (again) leaves the given region. Before this time point the trajectory can run either inside or outside of the region and it can reach the region and leave it again arbitrarily often. Region stability is thus characterized by the finiteness (NOT boundedness!) of the period of time that a trajectory can spend outside of the region.

The idea behind our result is based on two observations: there are exactly three basic situations that may be repeated finitely often in correct trajectories before stabilization, and it is possible to treat the three basic situations in a modular way. As a result, we can formulate three specific conditions, and we have shown that together they are necessary and sufficient for region stability. The possibility to use well-established state abstraction techniques gives rise to an interesting potential of practicality, as indicated by the experiments with a prototypical implementation of our tool.

References

- [1] A. Podelski and S. Wagner. Model checking of hybrid systems: From reachability towards stability. In J. P. Hespanha and A. Tiwari, eds., *Hybrid Systems: Computation and Control, 9th International Workshop, HSCC 2006*, Santa Barbara, CA, USA, March 2006, *LNCS 3927*, pp. 507–521. Sringer.
- [2] S. Wagner. A method and a tool for automatic verification of region stability for hybrid systems. Research Report MPI-I-2007-2-001, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, January 2007.

13.6 Formal Analysis of Business Processes

13.6.1 Formal semantics of Event-driven Process Chains

Investigator: Jörn Freiheit in collaboration with Ekkart Kindler (University Paderborn)

Event-driven Process Chains (EPCs) have been introduced in the early 90ties for modeling business processes and have been widely used since due to the incorporation into the ARIS tool set. For easing the modeling of business processes with EPCs, the informal semantics proposed for the OR-join and the XOR-join connectors are non-local. This non-locality results in severe problems when it comes to a formalization of the semantics of EPCs. Formalization, however, is crucial for formal analysis, verification and optimization of the business processes modeled with EPCs. To provide a formal semantics for EPCs we have developed a new algorithm to compute the state transition system of an EPC. The algorithm works for large EPCs as long as the resulting transition systems are small (where transition systems with millions of states and transitions are still considered to be small). The new technique for calculating the semantics of EPCs combines an explicitly forward construction of the transition system with a backward marking algorithm for checking the non-local constraints. Though this method does not always provide a result, it works for most practical examples and, in most cases, it is much faster than the until then best algorithm for computing the semantics of EPCs, which was based on symbolic model checking. The new algorithm has been implemented into an open source tool for the modeling and analysis of EPCs, called EPC Tools[1, 2].

References

- [1] N. Cuntz, J. Freiheit, and E. Kindler. On the semantics of EPCs: Faster calculation for EPCs with small state spaces. In M. Nuettgens and F. J. Rump, eds., *EPK 2005,*

Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten, Hamburg, Germany, 2005, pp. 7–23. Gesellschaft für Informatik.

- [2] C. Simon, J. Freiheit, and S. Olbrich. Using bpel processes defined by event-driven process chains. In M. Nüttgens, F. J. Rump, and J. Mendling, eds., *5. GI-Workshop "EPK 2006 - Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten"*, Vienna, Austria, 2006, pp. 121–135. CEUR Workshop Proceedings.

13.6.2 Formalization and analysis of rights delegation and revocation

Investigator: Jörn Freiheit in collaboration with Andreas Schaad (SAP Research)

Safety is a crucial issue for business processes. Access control to critical data and authorization to process certain tasks need to be managed. Users are often reluctant to introduce workflow management systems because of safety reasons. We have developed a formal model of delegation and revocation of rights and show how it can be easily combined with control flow models. Delegation and revocation of rights are main tasks for access control. Formal models are chosen because of their ability to be analyzed. This work is part of *R4eGov* (architecture for eGovernment), an EU-project within the 6th framework programme (see www.r4egov.info). R4eGov is bringing together public administrations and eGovernment experts from across the European Union in order to provide a closer collaboration of public administrations and to help them become more efficient. An appropriate access control ensures both that a task can be processed by an authorized principal and cannot be processed by a non-authorized one. However, the delegation and revocation of rights can be complex and formal methods are required to ensure the correctness of access control algorithms. We use a formal method, Petri nets, to model and analyze access right delegation and revocation schemes. A current trend of business processes is that different aspects, such as organizational, control and informational aspects, of the processes are modeled and implemented independently and are joined for workflow execution. We follow this approach and present a model of delegation and revocation that is independent of the control flow model of the business process. Hence, the control flow can be considered as a black box when defining the rights and their delegation and revocation. Additionally, the access control model can be considered as a black box when modeling the control flow. This distinction eases the modeling of the parts of the workflow and hence decreases modeling error rate. It is also possible to use different techniques for the modeling of the different aspects of the workflow. However, composing the models for control flow and access control leads to a combined model that can be executed and analyzed. Results are achieved by analyzing the access control model separately and then present some requirements that need to be observed when combining the model with a control flow model. Formal analysis of access control is a crucial task. Safety critical systems must ensure that no non-authorized principal gets access. However, because these systems can be very complex, a formal verification of correctness is often missing. Based on our model for rights delegation and revocation we show how properties can automatically be verified. The Petri net theory provides an abundance of analysis methods, most of them are implemented in CPNTools, which we use for automatic verification. CPNTools is based on the functional language ML, which supports formal reasoning and proof systems, such as HOL-ML. Using this tool we have analyzed several safety and liveness properties for a given initial configuration of rights appointments to principals and their delegation domain.

However, the model can be analyzed for an arbitrary configuration[1]. The model comprises different types of revocation, such as local, global, strong and weak revocation.

References

- [1] J. Freiheit. An initial step towards the formalisation of rights delegation and revocation using petri nets. In *9th International Conference on Enterprise Information Systems*, Funchal, Madeira, Portugal, 2007. accepted for publication.

13.6.3 Model-based representation of legal procedures

Investigators: Jörn Freiheit in collaboration with DFKI and Institute for Law and Informatics (Saarland University)

We have developed a concept that represents legal procedures in such a way as to make them more comprehensible[1, 2, 3]. This concept has been implemented into a demonstrator called Lexecute. Two steps were taken to formulate an adequate workflow model. The first step was to interview practitioners and consult legal texts. This step enabled the creation of semi-formal models in preparation for the second step, which was to verify the findings and formalize the processes. The workflow models resulting from this work provide new perspectives into justice, and reveal new potentials for modeling methods in the field of justice. In the eJustice project, an EU funded integrated project, judicial processes were modeled using Event-driven Process Chains, which then were transformed into Petri nets in order to overcome the problem of missing formal semantics. Petri nets were used for verification of judicial processes while Event-driven Process Chains were extended by additional information, such as legal basis for certain steps, deadlines, actors and organizational duties. Out of the model a web-based representation is automatically generated in order to provide an interface to judicial experts. This interface acts like a user-guide through the judicial procedure. Lexecute is a web-based demonstrator of this representation and have been applied successfully to several German and European judicial procedures.

References

- [1] J. Freiheit, M. Luuk, S. Münch, G. Sijanski, and F. Zangl. Lexecute: Visualisation and representation of legal procedures. *Digital Evidence Journal*, 3(1):17–27, May 2006.

- [2] J. Freiheit and F. Zangl. Model-based user-interface management for public services. In D. Remenyi, ed., *6th European Conference on e-Government*, Marburg, Germany, April 2006, pp. 141–151. Academic Conferences Limited.

- [3] J. Freiheit and F. Zangl. Model-based user-interface management for public services. In D. Remenyi, ed., *6th European Conference on e-Government*, Marburg, Germany, April 2006, pp. 141–151. Academic Conferences Limited.

13.7 Planning

13.7.1 Directed Model Checking

Investigator: Jörg Hoffmann

When looking for bugs in the state space of a transition system, it makes sense to search the states in a particular order, defined by a heuristic function estimating the "bug distance" of any state. The key question is how the heuristic function should be computed. The basic idea is to solve (find a bug in) an abstracted problem in every search state, and use the length of the abstract error path as the distance estimate. We have explored two alternative techniques of this kind, one [3, 2] inspired by techniques from AI Planning, and another one based on predicate abstraction and abstraction refinement [1]. Both techniques are implemented inside UPPAAL, and yield significant speedups, enabling us to find bugs in systems that could previously not be dealt with.

References

[1] J. Hoffmann, J. Smaus, A. Rybalchenko, S. Kupferschmid, and A. Podelski. Using predicate abstraction to generate heuristic functions in uppaal. In *MoChArt*, 2007. Accepted.

[2] Kupferschmid, K. Dräger, J. Hoffmann, B. Finkbeiner, H. Dierks, A. Podelski, and G. Behrmann. Uppaal/dmc – abstraction-based heuristics for directed model checking. In *TACAS*, 2007. Accepted.

- [3] S. Kupferschmid, J. Hoffmann, H. Dierks, and G. Behrmann. Adapting an ai planning heuristic for directed model checking. In A. Valmari, ed., *Model checking software, 13th International SPIN Workshop (SPIN-06)*, Vienna, Austria, 2006, *LNCS 3925*, pp. 35–52. Springer.

13.7.2 Problem Structure and DPLL Search

Investigator: Jörg Hoffmann

Modern DPLL-based procedures are extremely efficient on CNF formulas generated from practical applications such as AI Planning and Verification (i.e., Bounded Model Checking). Formulas with millions of variables can be solved, and even proved unsolvable, in a matter of minutes. This is in stark contrast to the worst-case complexity of the SAT problem. Clearly, practical applications exhibit certain kinds of structure that modern DPLL can exploit. The question is, can one find formal characterizations of what this "structure" is? We have explored this topic in the context of AI Planning. We have identified a high-level problem parameter that, as we proved in a comprehensive set of experiments, correlates with DPLL performance. We have analyzed synthetic examples were the amount of structure can be controlled, and we have proved the existence of doubly exponential gaps in DPLL best-case performance, between the two extremes (no structure/fully structured) [1].

References

[1] J. Hoffmann, C. Gomes, and B. Selman. Structure and problem hardness: Goal asymmetry and dpll proofs in sat-based planning. In D. Long and S. Smith, eds., *Proceedings of the 16th*

International Conference on Automated Planning and Scheduling (ICAPS-06), Ambleside, UK, 2006, pp. 284–293. kauf.

13.7.3 Probabilistic Planning

Investigator: Jörg Hoffmann

In many real-world planning situations – e.g., planning the actions for an autonomous robot – the initial state of the world is not known precisely, and the available actions may have different outcomes (e.g., they may fail with a certain likelihood). On the other hand, in recent years dramatic performance improvements were made for planning tools in a fully deterministic setting, where such uncertainty is not allowed. In our work, we have extended techniques from the deterministic setting to a probabilistic setting, where the initial "world" is a probability distribution over worlds, specified as a Bayesian Network, and every action specifies a probability distribution over its possible effects. We have developed a suitable representation of search states, and a domain-independent heuristic function. The resulting tool outperforms all other existing techniques in a significant range of benchmarks [1].

References

- [1] C. Domshlak and J. Hoffmann. Fast probabilistic planning through weighted model counting. In *16th International Conference on Automated Planning and Scheduling (ICAPS-06)*, The English Lake District, 2006, pp. 1–10. AAAI.

13.8 Academic Activities

13.8.1 Journal Positions

Jörg Hoffmann is on the editorial board of

- *Journal of Artificial Intelligence Research (JAIR).*

Stefan Ratschan is on the editorial board of

- *Mathematics in Computer Science.*

13.8.2 Conference and Workshop Positions

Membership in Program Committees

Peter Baumgartner:

- *Conference on Automated Deduction (CADE)*, Trustee of the steering committee, (elected 2003, re-elected 2006),
- *International Workshops on First-Order Theorem Proving (FTP)*, President of the steering committee (2003-2006),
- *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX)*, Member of the steering committee, (2003-2006),

- *International Joint Conference on Automated Reasoning (IJCAR)*, Member of the steering committee, (2004-2006),
- *Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2005)*, Koblenz, September 2005,
- *28th German Conference on Artificial Intelligence (KI-2005)*, Koblenz, September 2005,
- *Conference on Automated Deduction (CADE-20)*, Tallin, Estland, July 2005,
- *International Symposium for Methodologies on Intelligent Systems (ISMIS'05)*, Saratoga Springs, USA, 2005.

Jörg Hoffmann:

- *International Conference on Automated Planning and Scheduling (ICAPS 2005)*, Monterey, USA, June 2005,
- *The 20th National Conference on Artificial Intelligence (AAAI 2005)*, Pittsburgh, USA, July 2005.
- *Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005*, Edinburgh, Scotland, August 2005.
- *International Conference on Automated Planning and Scheduling (ICAPS 2006)*, English Lake District, UK, June 2006.
- *The 21st National Conference on Artificial Intelligence (AAAI 2006)*, Boston, USA, July 2006.
- *The 17th European Conference on Artificial Intelligence (ECAI 2006)*, Riva del Garda, Italy, August 2006.

Hans de Nivelle:

- *Workshop on Disproving, Non-Theorems, Non-Validity, Non-Provability*, Tallinn Estonia, July 2005,
- *Workshop on Disproving, Non-Theorems, Non-Validity, Non-Provability*, Seattle, Washington, August 2006,
- *The sixth International Workshop on the Implementation of Logics*, Phnom Penh, Cambodia, November 2006.

Stafan Ratschan:

- *International Conference on Mathematical Aspects of Computer and Information Sciences*, Beijing, China, 2006,
- *8-th Asian Symposium on Computer Mathematics (ASCM)*, Seoul, 2005,
- *Workshop on Interval Analysis and Constraint Propagation for Applications*, Sitges, Spain, 2005.
- *Workshop on Quantification in Constraint Programming*, Sitges, Spain, 2005.

Membership in Organizing Committees

Peter Baumgartner:

- *Workshop on Disproving in conjunction with the 20th International Conference on Automated Deduction (CADE-20)*, Tallin, Estonia, 2005,
- *Workshop at 28th German Conference on Artificial Intelligence KI-2005)*, chair, Koblenz, Germany, 2005,
- *Dagstuhl-Seminar Deduction and Applications*, Schloss Dagstuhl, Germany, October 2005.

Stafan Ratschan:

- *Workshop on Interval Analysis and Constraint Propagation for Applications*, Sitges, Spain, 2005.

13.8.3 Invited Talks and Tutorials

Andrey Rybalchenko:

- *Automated termination proofs for systems code*, Invited talk,
  - 8th International Workshop on Verification of Infinite-State Systems, Bonn, Germany, August 2006.
  - Alpine Verification Meeting 2006, Monte Verita, Switzerland, May 2006.
  - Seminar on Software Verification: Infinite-State Model Checking and Static Program Analysis. Dagstuhl, February 2006.
  - Model Checking Day at Theoretical Computer Science Chair, TU München. Dec 2005.
  - Seminar at L.I.A.F.A., Paris 7 University, November 2005.
  - Fundamental Computer Science F.N.R.S. Contact Group, Montefiore Institute, University of Liège, October 2005.
- *Transition predicate abstraction and fair termination*, Invited talk, TRESOR seminar at EPFL, Lausanne, April 2005.

Jörg Hoffmann:

- *Local Search and Problem Structure in AI Planning*, Invited talk, Principles and Practice of Constraint Programming (CP), Barcelona, Spain, October 2005.

Hans de Nivelle:

- *Introduction to Automated Reasoning* at the 18th European Summer School in Logic, Language and Information (ESSLLI 2006), August 2006.

13.9 Teaching Activities

Courses:

Introduction to Proof Theory. lecture for graduate students, 2005 (H. de Nivelle, R. Piskac)

13.10 Dissertations, Habilitations, Offers, Awards

13.10.1 Dissertations

- Andrey Rybalchenko: *Transition Invariants for Automated Temporal Verification*, summa cum laude, June 2005
- Yevgeny Kazakov: *Saturation-Based Decision Procedures for Extensions of the Guarded Fragment*, magma cum laude, March 2006

13.10.2 Offers for Faculty Positions

- Andreas Podelski, accepted C4 professorship for software technology, University Freiburg, 2006.

13.10.3 Awards

- Andrey Rybalchenko, 2006 Otto Hahn Medal awarded by the Max Planck Society
- Jörg Hoffmann, Best paper prize 2005 at IJCAII-Journal of Artificial Intelligence, together with Bernhard Nebel (University of Freiburg).

13.11 Grants and Cooperations

AVACS – Automatic Verification and Analysis of Complex Systems

AVACS is a transregional collaborative research center funded by the German Research Foundation (DFG). The center addresses the rigorous mathematical analysis of models of complex safety critical computerized systems, 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 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. Within this research center, our group is working on the development of new decision procedures and constraint solvers, on the integration of deductive approaches and on improving abstraction techniques for model checking real-time, hybrid or dynamic communicating systems.

- Starting date: January 2005
- Duration: 12 years
- Funding: DFG Transregional Collaborative Research Center
- Staff at MPI: Andreas Podelski, Stefan Ratschan, Viorica Sofronie-Stokkermans, Uwe Waldmann, Patrick Maier, Jörg Hoffmann, Ina Schaefer, Silke Wagner, Swen Jacobs, She Zhikun
- Partners: Universities of Freiburg, Oldenburg, Saarbrücken

Verisoft

Verisoft is a collaborative project of several academic and industrial partners, funded by the German federal ministry of education and research. The general goal of the project is to formally verify all layers (hardware, operating system, software) of safety- or security-critical systems such as drive-by-wire controllers in automobiles or cryptographic smart-cards. Our focus is on the development of automatic tools (theorem provers, model checkers) for software verification and on their integration into interactive verification environments.

- Starting date: July 2003
- Duration: 8 years
- Funding: BMBF
- Staff at MPI: Andreas Podelski, Uwe Waldmann, Carsten Ihlemann, Virgile Prevosto, Patrick Maier, Thomas Hillenbrand, Mohamed Nassim Seghir, Stefan Maus, Ina Schaefer, Thomas Wies
- Partners: BMW, Infineon, T-Systems, DFKI, Universities of Darmstadt, Karlsruhe, Koblenz, München

R4eGov – Architecture for eGovernment

Integrated research and development project, supported by the European Commission's FP6 IST program, work on formalisation and verification of workflows in European public processes.

- Starting date: March 2006
- Duration: 3 years
- Staff at MPI: Andreas Podelski, Jörn Freiheit
- Partners: SAP, Thales, Unisys, DFKI, ...

eJustice

Integrated research and development project, supported by the European Commission's FP6 IST program, work on formalisation and verification of workflows in transnational juridicial processes.

- Starting date: March 2004
- Duration: 2 years
- Staff at MPI: Andreas Podelski, Jörn Freiheit
- Partners: SAP, Thales, Dept. of Law at University of Saarland, DFKI, ...

Abstraction for Verification

Foundational Research on the automated refinement of abstraction in software model checking methods.

- Starting date: January 2002

- Duration: 3 years
- Funding: DAAD (in the VIGONI programme)
- Staff at MPI: Andreas Podelski, Andrey Rybalchenko
- Partners: Roberto Giacobazzi, University of Verona, Italy

OntoNat – Ontological Reasoning for Natural Language Understanding

The goal of the project is to develop techniques and systems for information retrieval and question answering purposes. Our focus is on a hydrid approach combining state-of-the-art techniques from knowledge representation, computational linguistics and automated reasoning.

- Starting date: October 2003
- Duration: 2 years
- Funding: MPI
- Staff at MPI: Peter Baumgartner, Fabian Suchanek
- Partners: University of Saarland, Department of Computational Linguistics and Phonetics (Prof. Pinkal).

13.12 Publications

Books

[1] W. Ahrendt, P. Baumgartner, and H. de Nivelle, eds. *IJCAR'06 Workshop, Disproving'06: Non-Theorems, Non-Validity, Non-Provability*, Seattle, USA, 2006. The 2006 Federated Logic Conference.

Journal articles

[1] P. Baumgartner, A. Fuchs, and C. Tinelli. Implementing the model evolution calculus. *International Journal on Artificial Intelligence Tools*, 15(1):21–52, 2006.

[2] P. Baumgartner, U. Furbach, and A. Yahya. Automated reasoning, knowledge representation and management. *KI - Künstliche Intelligenz*, 1:5–11, 2005.

[3] B. Blanchet and A. Podelski. Verification of cryptographic protocols: Tagging enforces termination. *Theoretical Computer Science*, 333(1-2):67–90, March 2005.

[4] D. Delahaye, M. Jaume, and V. Prevosto. Coq, un outil pour l'enseignement. *Technique et Science Informatiques*, 24(9):1139–1160, 2005.

[5] J. Freiheit, M. Luuk, S. Münch, G. Sijanski, and F. Zangl. Lexecute: Visualisation and representation of legal procedures. *Digital Evidence Journal*, 3(1):17–27, May 2006.

[6] H. Ganzinger and J. Stuber. Superposition with equivalence reasoning and delayed clause normal form transformation. *Information and Computation*, 199(1-2):3–23, 2005.

[7] S. Goncharov, V. Harizanov, J. F. Knight, A. Morozov, and A. Romina. On automorphic tuples of elements in computable models. *Siberian Mathematical Journal*, 46(3):405–412, 2005.

[8] J. Hoffmann. The deterministic part of ipc-4: An overview. *Journal of Artificial Intelligence Research*, 24:519 – 579, October 2005.

[9] J. Hoffmann. In defense of pddl axioms. *Artificial Intelligence*, 168(1-2):38–69, 2005.

[10] J. Hoffmann. Where ignoring delete lists works: Local search topology in planning benchmarks. *Journal of Artificial Intelligence Research*, 24:685–758, November 2005.

[11] J. Hoffmann and R. I. Brafman. Conformant planning via heuristic forward search: A new approach. *Artificial Intelligence*, 170:507–541, 2006.

[12] K. Jensen and A. Podelski. Tools and algorithms for the construction and analysis of systems. *International Journal on Software Tools for Technology Transfer*, 8(3):177–179, 2006.

[13] K. Korovin and A. Voronkov. Knuth-Bendix constraint solving is NP-complete. *ACM Transactions on Computational Logic*, 6(2):361–388, 2005.

[14] H. de Nivelle. Translation of resolution proofs into short first-order proofs without choice axioms. *Information and Computation*, 199(1):24–54, April 2005.

[15] H. de Nivelle and S. Demri. Deciding regular grammar logics with converse through first-order logic. *Journal of Logic, Language and Information*, 14(3):289–329, June 2005.

[16] S. Ratschan. Efficient solving of quantified inequality constraints over the real numbers. *ACM Transactions on Computational Logic*, 7(4):723–748, 2006.

[17] S. Ratschan and Z. She. Safety verification of hybrid systems by constraint propagation based abstraction refinement. *ACM Transactions in Embedded Computing Systems (TECS)*, 6(1):8, 2007.

Conference articles

[1] P. Baumgartner and U. Furbach. Living books, automated deduction and other strange things. In D. Hutter and W. Stephan, eds., *Mechanizing Mathematical Reasoning: Techniques, Tools and Applications - Essays in honour of Jörg H. Siekmann*, *LNAI 2605*, pp. 255–274. Springer, New York, USA, 2005.

[2] P. Baumgartner, U. Furbach, M. Gross-Hardt, and T. Kleemann. Optimizing the Evaluation of XPath Using Description Logics. In D. Seipel, M. Hanus, U. Geske, and O. Bartenstein, eds., *Applications of Declarative Programming and Knowledge Management: 15th International Conference on Applications of Declarative Programming and Knowledge Management, INAP 2004, and 18th Workshop on Logic Programming, WLP 2004*, Potsdam, Germany, January 2005, *LNAI 3392*, pp. 1–15. Springer.

[3] P. Baumgartner and C. Tinelli. The model evolution calculus with equality. In R. Nieuwenhuis, ed., *Automated deduction - CADE-20, 20th International Conference on Automated Deduction*, Tallinn, Estonia, 2005, *LNAI 3632*, pp. 392–408. Springer.

[4] W. Charatonik, L. Georgieva, and P. Maier. Bounded model checking of pointer programs. In L. Ong, ed., *Computer Science Logic; 19th International Workshop, CSL 2005; 14th Annual Conference of the EACSL*, Oxford, UK, August 2005, *LNCS 3634*, pp. 397–412. Springer.

[5] B. Cook, A. Podelski, and A. Rybalchenko. Abstraction-refinement for termination. In C. Hankin and I. Siveroni, eds., *Static analysis, 12th International Symposium, SAS 2005*, London, UK, September 2005, *LNCS 3672*, pp. 87–101. Springer.

[6] B. Cook, A. Podelski, and A. Rybalchenko. Termination proofs for systems code. In *PLDI 2006, Proceedings of the ACM SIGPLAN 2006 Conference on Programming Language Design and Implementation*, Ottawa, Ontario, Canada, 2006, *ACM SIGPLAN Notices*, vol. 41, pp. 415–426. ACM.

[7] B. Cook, A. Podelski, and A. Rybalchenko. Terminator: Beyond safety. In T. Ball and R. B. Jones, eds., *Computer aided verification, 18th International Conference, CAV 2006*, Seattle, WA, USA, 2006, *LNCS 4144*, pp. 415–418. Springer.

[8] N. Cuntz, J. Freiheit, and E. Kindler. On the semantics of EPCs: Faster calculation for EPCs with small state spaces. In M. Nuettgens and F. J. Rump, eds., *EPK 2005, Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten*, Hamburg, Germany, 2005, pp. 7–23. Gesellschaft für Informatik.

[9] W. Damm, G. Pinto, and S. Ratschan. Guaranteed termination in the verification of ltl properties of non-linear robust discrete time hybrid systems. In D. A. Peled and Y.-K. Tsay, eds., *Automated technology for verification and analysis, Third International Symposium, ATVA 2005*, Taipei, Taiwan, 2005, *LNCS 3707*, pp. 99–113. Springer.

[10] M. Daum, S. Maus, N. Schirmer, and M. N. Seghir. Integration of a software model checker into isabelle. In G. Sutcliffe and A. Voronkov, eds., *Logic for Programming, Artificial Intelligence, and Reasoning: 12th International Conference, LPAR 2005*, Montego Bay, Jamaica, October 2005, *LNAI 3835*, pp. 381–395. Springer.

[11] C. Domshlak and J. Hoffmann. Fast probabilistic planning through weighted model counting. In *16th International Conference on Automated Planning and Scheduling (ICAPS-06)*, The English Lake District, 2006, pp. 1–10. AAAI.

[12] K. Dräge, B. Finkbeiner, and A. Podelski. Directed model checking with distance-preserving abstractions. In A. Valmari, ed., *Model checking software, 13th International SPIN Workshop*, Vienna, Austria, 2006, *LNCS 3925*, pp. 19–34. Springer.

[13] J. Freiheit. An initial step towards the formalisation of rights delegation and revocation using petri nets. In *9th International Conference on Enterprise Information Systems*, Funchal, Madeira, Portugal, 2007. accepted for publication.

[14] J. Freiheit and F. Zangl. Model-based user-interface management for public services. In D. Remenyi, ed., *6th European Conference on e-Government*, Marburg, Germany, April 2006, pp. 141–151. Academic Conferences Limited.

[15] J. Freiheit and F. Zangl. Model-based user-interface management for public services. In D. Remenyi, ed., *6th European Conference on e-Government*, Marburg, Germany, April 2006, pp. 141–151. Academic Conferences Limited.

[16] 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.

[17] L. Georgieva and P. Maier. Description logics for shape analysis. In B. K. Aichernig and B. Beckert, eds., *Third IEEE International Conference on Software Engineering and Formal Methods (SEFM 2005)*, Koblenz, Germany, September 2005, pp. 321–330. IEEE.

[18] J. Hoenicke and P. Maier. Model-checking of specifications integrating processes, data and time. In J. Fitzgerald, I. J. Hayes, and A. Tarlecki, eds., *FM 2005: Formal Methods; International Symposium of Formal Methods Europe*, Newcastle, UK, July 2005, *LNCS 3582*, pp. 465–480. Springer.

[19] J. Hoffmann. A covering problem for hypercubes. In L. Kaelbling, ed., *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, New York, USA, 2005, pp. 579–580. Morgan Kaufmann.

[20] J. Hoffmann. Friends or foes? an ai planning perspective on abstraction and search. In *16th International Conference on Automated Planning and Scheduling (ICAPS-06)*, The English Lake District, 2006, pp. 1–10. AAAI.

[21] J. Hoffmann. Structure and problem hardness: Goal asymmetry and dpll proofs in sat-based planning. In *16th International Conference on Automated Planning and Scheduling (ICAPS-06)*, The English Lake District, 2006, pp. 1–10. AAAI.

[22] J. Hoffmann and R. Brafman. Contingent planning via heuristic forward search with implicit belief states. In S. Biundo, K. Meyers, and K. Rajan, eds., *15th International Conference on Automated Planning and Scheduling*, Monterey, USA, June 2005, pp. 71–80. AAAI.

[23] S. Kupferschmid, J. Hoffmann, H. Dierks, and G. Behrmann. Adapting an ai planning heuristic for directed model checking. In A. Valmari, ed., *Model checking software, 13th International SPIN Workshop (SPIN-06)*, Vienna, Austria, 2006, *LNCS 3925*, pp. 35–52. Springer.

[24] N. Lilith, J. Billington, and J. Freiheit. Approximate closed-form aggregation of a fork-join structure in generalised stochastic petri nets. In L. Lenzini and R. L. Cruz, eds., *Proceedings of the 1st International Conference on Performance Evaluation Methodolgies and Tools, VALUETOOLS 2006*, Pisa, Italy, 2006, *ACM International Conference Proceeding Series*, vol. 180, pp. 1–10. ACM.

[25] A. Malkis, A. Podelski, and A. Rybalchenko. Thread-modular verification is cartesian abstract interpretation. In K. Barkaoui, A. Cavalcanti, and A. Cerone, eds., *Theoretical aspects of computing - ICTAC 2006, third International Colloquium*, Tunis, Tunisia, 2006, *LNCS 4281*, pp. 183–197. Springer.

[26] H. de Nivelle, P. Baumgartner, A. Fuchs, and C. Tinelli. Computing finite models by reduction to function-free clause logic. In W. Ahrendt, P. Baumgartner, and H. de Nivelle, eds., *IJCAR'06 Workshop, Disproving'06: Non-Theorems, Non-Validity, Non-Provability*, Seattle, USA, August 2006, pp. 82–95. The 2006 Federated Logic Conference.

[27] H. de Nivelle and J. Meng. Geometric resolution: A proof procedure based on finite model search. In U. Furbach and N. Shankar, eds., *Automated reasoning, Third International Joint Conference, IJCAR 2006*, Seattle, WA, USA, August 2006, *LNAI 4130*, pp. 303–317. Springer.

[28] H. de Nivelle and R. Piskac. Verification of an off-line checker for priority queues. In B. K. Aichernig and B. Beckert, eds., *Third IEEE International Conference on Software Engineering and Formal Methods (SEFM 2005)*, Koblenz, 2005, pp. 210–219. IEEE.

[29] A. Pnueli, A. Podelski, and A. Rybalchenko. Separating fairness and well-foundedness for the analysis of fair discrete systems. In N. Halbwachs and L. Zuck, eds., *Tools and Algorithms for the Construction and Analysis of Systems: 11th International Conference, TACAS 2005, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2005*, Edinburgh, UK, April 2005, *LNCS 3440*, pp. 124–139. Springer.

[30] A. Podelski and A. Rybalchenko. Transition predicate abstraction and fair termination. In J. Palsberg and M. Abadi, eds., *Proceedings of the 32nd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 2005*, Long Beach, CA, USA, 2005, pp. 124–139. ACM.

[31] A. Podelski, I. Schaefer, and S. Wagner. Summaries for while programs with recursion. In M. Sagiv, ed., *Programming Languages and Systems: 14th European Symposium on Programming, ESOP 2005, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2005*, Edinburgh, UK, April 2005, *LNCS 3444*, pp. 94–107. Springer.

[32] A. Podelski and S. Wagner. Model checking of hybrid systems: From reachability towards stability. In J. P. Hespanha and A. Tiwari, eds., *Hybrid Systems: Computation and Control, 9th International Workshop, HSCC 2006*, Santa Barbara, CA, USA, March 2006, *LNCS 3927*, pp. 507–521. Sringer.

[33] A. Podelski and T. Wies. Boolean heaps. In C. Hankin and I. Siveroni, eds., *Static analysis, 12th International Symposium, SAS 2005*, London, UK, September 2005, *LNCS 3672*, pp. 268–283. Springer.

[34] V. Prevosto. Certified mathematical hierarchies: the FoCal system. In T. Coquand, H. Lombardi, and M.-F. Roy, eds., *Proceedings of the MAP (Mathematics, Algorithms, Proofs) Workshop*, IBFI Schloß Dagstuhl, Germany, 2006, *Dagstuhl Seminar Proceedings*, vol. 05021. IBFI.

[35] V. Prevosto and S. Boulmé. Proof contexts with late binding. In P. Urzyczyn, ed., *Typed Lambda Calculi and Applications: 7th International Conference, TLCA 2005*, Nara, Japan, April 2005, *LNCS 3461*, pp. 325–339. Springer. To appear.

[36] S. Ratschan. Solving undecidable problems in the theory of real numbers and hybrid systems. In A. Dolzmann, A. Seidl, and T. Sturm, eds., *Algorithmic Algebra and Logic; Conference in Honor of the 60th Birthday of Volker Weispfenning*, Passau, Germany, 2005, pp. 213–216. Books on Demand GmbH.

[37] S. Ratschan and Z. She. Safety verification of hybrid systems by constraint propagation based abstraction refinement. In M. Morari and L. Thiele, eds., *Hybrid Systems: Computation and Control: 8th International Workshop, HSCC 2005*, Zürich, Schweiz, 2005, *LNCS 3414*, pp. 573–589. Springer.

[38] S. Ratschan and Z. She. Constraints for continuous reachability in the verification of hybrid systems. In J. Calmet, T. Ida, and D. Wang, eds., *Artificial Intelligence and Symbolic Computation, 8th International Conference, AISC 2006*, Beijing, China, September 2006, *LNAI 4120*, pp. 196–210. Springer. This work was partly supported by the German Research Council (DFG) as part of the Transregional Collaborative Research Center "Automatic Verification and Analysis of Complex Systems" (SFB/TR 14 AVACS). See www.avacs.org for more information.

[39] S. Ratschan and J.-G. Smaus. Verification-integrated falsification of non-deterministic hybrid systems. In *2nd IFAC Conference on Analysis and Design of Hybrid Systems*, Alghero, Sardinia, November 2006. Elsevier.

[40] A. Rybalchenko. Model checking duration calculus: a practical approach. In *Theoretical Aspects of Computing - ICTAC 2006, Third International Colloquium*, Tunis, Tunisia, 2006, *LNCS 4281*, pp. 332–346. Springer.

[41] A. Rybalchenko. Armc: the logical choice for software model checking with abstraction refinement. In *9th International Symposium on Practical Aspects of Declarative Languages (PADL 2007)*, Nice, France, 2007, *LNCS 4281*, pp. 245–259. Springer.

[42] A. Rybalchenko. Invariant synthesis for combined theories. In A. Podelski, ed., *8th International Conference on Verification, Model Checking, and Abstract Interpretation (VMCAI 2007)*, Nice, France, 2007, *LNCS 4349*, pp. 0–0. Springer.

[43] A. Rybalchenko. Path invariants. In *ACM SIGPLAN 2007 Conference on Programming Language Design and Implementation*, San Diego, USA, 2007, pp. 0–0. ACM.

[44] A. Rybalchenko. Proving that programs eventually do something good. In *34th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL 2007)*, Nice, France, 2007, pp. 265–276. ACM.

[45] A. Rybalchenko. Proving thread termination. In *ACM SIGPLAN 2007 Conference on Programming Language Design and Implementation*, San Diego, USA, 2007, pp. 0–0. ACM.

[46] Z. She. Providing a basin of attraction to a target region by computation of lyapunov-like functions. In *4th IEEE International Conference on Computational Cybernetics*, Tallinn, Estonia, August 2006, pp. 245–249. IEEE.

[47] Z. She, B. Xia, and R. Xiao. A semi-algebraic approach for the computation of lyapunov functions. In *2th IASTED International Conference on COMPUTATIONAL INTELLIGENCE*, San Francisco, CA, USA, 2006. ACTA Press.

[48] C. Simon, J. Freiheit, and S. Olbrich. Using bpel processes defined by event-driven process chains. In M. Nüttgens, F. J. Rump, and J. Mendling, eds., *5. GI-Workshop "EPK 2006 - Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten"*, Vienna, Austria, 2006, pp. 121–135. CEUR Workshop Proceedings.

[49] S. Wagner. A sound and complete proof rule for region stability of hybrid systems. In A. Bemporad, A. Bicchi, and G. Buttazzo, eds., *Proceedings of the 10th Conference on HYBRID SYSTEMS: COMPUTATION AND CONTROL*, Heidelberg, April 2007, *LNCS 4416*, pp. 750–753. Springer.

[50] T. Wies, V. Kuncak, P. Lam, A. Podelski, and M. C. Rinard. Field constraint analysis. In E. A. Emerson and K. S. Namjoshi, eds., *Verification, Model Checking, and Abstract Interpretation, 7th International Conference, VMCAI 2006*, Charleston, SC, USA, January 2006, *LNCS 3855*, pp. 157–173. Springer.

[51] C. Yuan, J. Billington, and J. Freiheit. An abstract model of routing in mobile ad hoc networks. In *Sixth Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools*, Aarhus, Denmark, October 2005, pp. 137–156. DAIMI.

Theses

[1] R. Dimitrova. Model checking with abstraction refinement for well-structured systems. Masters thesis, Universität des Saarlandes, June 2006.

[2] Y. Kazakov. *Saturation-Based Decision Procedures For Extensions Of The Guarded Fragment.* Phd thesis, Universität des Saarlandes, March 2006. Magna Cum Laude.

[3] R. Piskac. Formal correctness of result checking for priority queues. Masters thesis, Universität des Saarlandes, February 2005.

[4] A. Rybalchenko. *Transition Invariants for Automated Temporal Verification.* PhD thesis, Universität des Saarlandes, June 2005.

[5] M. Schäf. Abstrakte Übergangsrelationen als Mittel zur Verifikation von Programmeigenschaften. Masters thesis, Universität des Saarlandes, April 2006.

[6] F. M. Suchanek. Ontological reasoning for natural language understanding. Masters thesis, Universität des Saarlandes, 2005.

14 The Computational Biology and Applied Algorithmics Group (D3)

14.1 Personnel

Director
Prof. Dr. Thomas Lengauer, PhD

System Administration
Dr. Joachim Büch

Researchers
Dr. Mario Albrecht
Dr. Iris Antes
Dr. Francisco Silva Domingues
Dr. Andreas Kämper (–January 2007)
Dr. Gabriele Mayr
Dr. Jörg Rahnenführer (–March 2007)
Dr. Ingolf Sommer

PhD Students
Adrian Alexa
André Altmann (November 2005–)
Yassen Assenov (April 2007–)
Christoph Bock
Jasmina Bogojeska (January 2007–)
Konstantin Halachev (August 2006–)
Christoph Hartmann
Lars Kunert
Jochen Maydt
Fidel Ramírez
Kirsten Roomp
Oliver Sander
Andreas Schlicker (November 2005–)
Tobias Sing (–April 2007)
Andreas Steffen
Priti Talwar
Alexander Thielen (March 2006–)

Laura Tolosi (May 2006–)
Hongbo Zhu

Short-term Employees
Hagen Blankenburg (April 2007–)
Oliver Mueller (October 2006–)
Nils Weinhold (December 2006–)

Secretary
Ruth Schneppen-Christmann

14.2 Visitors

In the time period from March 2005 to March 2007, the following researchers visited our group:

Csaba Pal	27.04.2005	EMBL Heidelberg
Gabriele Weiler	15.06.2005	DKFZ Heidelberg
Silvio Tosatto	01.12.2005	University of Padua, Italy
Benno Schwikowski	22.11.2005	Institut Pasteur, Paris
Martin Ginkel	15.12.2005	Universität Magdeburg
Melissa Cline	20.03.2006	Institut Pasteur, Paris
Peter Revesz	07.04.2006	University of Nebraska, USA
Brian Chen	09.05.2006	Rice University Houston Texas
Wolfgang Schulz		Heinrich-Heine-Universität Düsseldorf
Torsten Schwede	10.10.2006	Universität Basel
Victor Neduva	25.10.2006	EMBL Heidelberg
Ranjit Bahadur	03.11.2006	Institut de Biochimie et Biosphysique, Paris
Martin Zacharias	07.12.2006	International University Bremen
Carsten Carlberg	28.11.2006	University of Luxemburg
Nils Weskamp	01.02.2007	Universität Marburg
Roman Thomas	14.02.2007	MPI für Neurologische Forschung Köln
Christian Weichenberger	22.02.2007	University of Salzburg, Austria

14.3 Group Organization

All group members except the director, the secretary, and the systems administrator have temporary positions. The group used to be without any internal hierarchy, until in November 2006 an departmental research group on *Molecular Networks in Medical Bioinformatics* has been established that is directed by Mario Albrecht and currently comprises five researchers

(including PhD students and excluding the group leader), two (masters) students and one guest. The group members report to the group leader on scientific matters and to the director on all other matters.

All other scientists including the doctoral students report directly to the director on all matters. 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 group. Bachelor and masters students are generally advised by the scientists.

The group 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 from an external guest) and for discussions and announcements of general group interest. Recently, student thesis talks have become so frequent that they are often scheduled separately. The group maintains a set of only internally accessible web pages that collect information relevant for the group such as central locally maintained databases, the status of student projects, seminar schedules and the like.

14.4 HIV Bioinformatics

14.4.1 Resistance Analysis

Investigators: André Altmann, Tobias Sing, and Joachim Büch

Today, 21 approved antiretroviral agents are available to treat HIV infections. The drugs are grouped into 5 classes according to their molecular target and their mechanism of action. Entry inhibitors (EIs) are applied to prevent the viral particle from entering the host cell by binding to the viral transmembrane protein gp41 (for details see 14.4.2). Currently there is only one approved EI, but several more (about 10) are under investigation. Nucleoside (and nucleotide) reverse transcriptase (RT) inhibitors (NRTIs) interrupt the process of reverse transcription by acting as terminators of DNA chain elongation. Their mechanism of action differs from non-nucleoside RT inhibitors (NNRTIs) that bind to the active side of the viral enzyme RT. The fourth group are integrase inhibitors (IIs) that prevent the viral enzyme integrase from integrating the (reverse-)transcribed viral genetic material into the genome of an infected cell. Today, two IIs are under investigation, but none is approved. The last group are protease inhibitors (PIs). PIs inhibit the viral enzyme protease (PR) that is involved into maturation of new infectious viral particles.

The need for the large number of pharmaceuticals that attack different stages of the viral life cycle arose due to the high mutation rate of HIV. The process of reverse transcription is highly error-prone due to lack of a proof-reading mechanism. The mutation rate of HIV-1 is estimated to be in the range of 10^{-4} to 10^{-6} substitutions per base pair per replication cycle. In the presence of antiretroviral drugs viral variants harboring so-called *drug resistance* mutations can replicate better than the wild-type. Thus, drug resistant mutants will outcompete the wild type and render the therapy useless. Hence, to date mono-therapies are replaced by combination therapies. The goal of this so-called highly active antiretroviral therapy (HAART) is to combine three or more drugs, typically from at least two different classes, in order to suppress HIV replication and to prevent progression to AIDS and death.

From genotype to phenotype

Genotypic assays are the standard methods for guiding treatment selection for patients infected with HIV-1. However, due to dependencies between resistance mutations interpretation of the genotypic information is a challenging task. Today, algorithms that apply hand-crafted expert rules to analyze the resistance mutations provide so-called genotypic susceptibility scores for every drug. Another method is to measure the susceptibility of a virus against a specific antiretroviral agent in an laboratory assay. Compared to genotyping this method is time- and cost-intense, but provides an easy to interpret result: the *resistance factor*. Thus, in earlier work we introduced the system GENO2PHENO [3] (reviewed in [9]) that uses support vector regression to compute phenotypic resistance from the viral genotype. This obtained phenotypic resistance scores can be interpreted exactly like the results obtained from the laboratory assay.

The development of resistance involves the stochastic accumulation of mutations in the viral genome along certain mutational pathways. In earlier work we introduced the notion of the mutagenetic trees, a family of graphical models designed to represent such genetic accumulation processes [5]. In order to improve the quality of the geno2pheno system we investigated the use of a Fisher kernel derived from mutagenetic tree mixtures [11]. The introduced kernel exploits the evolutionary structure in genotype-phenotype prediction. Our results showed a significant improvement in performance across 17 anti HIV drugs compared to the support vector regression.

Therapy Optimization

The rationale for HAART is to maximally suppress virus replication and to avoid (or at least delay) the development of drug resistance. However, with the large number of drugs available and novel drugs being approved almost every year, it becomes increasingly difficult for the treating physician to select an optimal drug combination. To date, most methods predict virological response to therapy based on the baseline genotype and the compounds in the applied combination. In [8] neural networks were used to predict the change in viral load (copies of viral RNA/ml blood plasma) after the onset of a new therapy. We used a new way of analyzing so-called treatment change episodes (TCEs). A TCE consists of a viral genotype, the compounds of the selected treatment, and a binary outcome indicating success or failure of the regimen. We compared different encodings of the viral genotype and antiretroviral regimen including phenotypic and evolutionary information, namely predicted phenotypic drug resistance (GENO2PHENO), activity of the regimen estimated from sequence space search [4], the genetic barrier to drug resistance [6], and the genetic progression score (see also: 14.7.2). These derived feature were evaluated in the context of different statistical learning procedures applied to the binary classification task of predicting virological response. Classifier performance was evaluated on about 6,300 observed TCEs from the Stanford HIV Drug Resistance Database and a large US clinic-based patient population. We found that the genetic barrier to drug resistance and predicted phenotypic drug resistance are the best encodings across all datasets and statistical learning methods examined [1].

Based on these results we implemented a prototypical therapy ranker: THEO (THErapy Optimizer). THEO takes as input the sequences of RT and PR of the viral genome, and

computes the probability of success for a predefined set of combination therapies. These success scores are the basis of the generated ranking.

Within the EU project Eu*Resist* databases from three European institutes (Germany: AREVIR, Italy: ARCA, Sweden: Karolinska) are combined to enlarge the number of available clinical data. Our part in the project involves building a prediction engine similar to THEO. This engine and the engines developed by other Eu*Resist* partners will be combined to optimize the prediction quality. The consortium's aim is to provide a free to access prediction system to predict response to highly active antiretroviral therapy [12] (paper submitted to the IST-Africa 2007 Conference).

Clinical validation

To date, several tools providing support for interpreting the genotypic information exist. GENO2PHENO and THEO are such tools. In order to prove clinical relevance, a statistical comparison to state-of-the-art (rule-based) interpretation systems is called for. We evaluated THEO and GENO2PHENO on a large set of retrospective clinical data obtained from the Eu*Resist* database. The systems were compared in a regression setting, where the change in viral load after start of the treatment had to be predicted, and in a classification setting, where response to antiretroviral therapy was dichotomized. Obtained results were compared to three rule-based algorithms: HIVDB [10], ANRSV2006.07, and REGA [7]. We found that treatment selection guided by phenotypic susceptibility scores (as given by GENO2PHENO) performs as well as with genotypic susceptibility scores (as provided by the rule-based algorithms). However, THEO outperforms both approaches if response is dichotomized to success and failure [2].

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, T. Sing, N. Beerenwinkel, H. Walter, R. Kaiser, and T. Lengauer. Validation of geno2pheno and THEO on a large independent clinical dataset. In *Proceedings of the 5th European HIV Drug Resistance Workshop*, Paris, France, 2007, HIV Medicine. European AIDS Clinical Society.

[3] 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.

[4] N. Beerenwinkel, T. Lengauer, M. Däumer, R. Kaiser, H. Walter, K. Korn, D. Hoffmann, and J. Selbig. Methods for optimizing antiviral combination therapies. *Bioinformatics*, 19(Supplement):i16–i25, 2003.

- [5] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer. Learning multiple evolutionary pathways from cross-sectional data. *Journal of Computational Biology*, 12(6):584–598, 2005.

[6] N. Beerenwinkel, T. Sing, M. Däumer, R. Kaiser, and T. Lengauer. Computing the genetic barrier. In C. Boucher, J. Mellors, B. Larder, and D. Richman, eds., *XII International HIV Drug resistance Workshop*, Tenerife, Spain, 2004, *Antiviral Therapy*, vol. 9, pp. S125–S125. International Medical Press.

[7] K. V. Laethem, A. D. Luca, A. Antinori, A. Cingolani, C. F. Perna, and A.-M. Vandamme. A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in hiv-1-infected patients. *Antiviral Therapy*, 7(2):123–129, Jun 2002.

[8] B. Larder, D. Wang, A. Revell, J. Montaner, R. Harrigan, F. D. Eolf, J. Lange, S. Wegner, L. Ruiz, M. Jesus, Perez-Elia, S. Emery, J. Gatell, A. D. Monforte, C. Torti, M. Zazzi, and C. Lane. The development of artificial neural networks to predict virological response to combination hiv therapy. *Antiviral Therapy*, 12(1):15–24, Feb 2007.

• [9] T. Lengauer and T. Sing. Bioinformatics-assisted anti-HIV therapy. *Nature Reviews Microbiology*, 4:790–797, October 2006.

[10] R. W. Shafer, D. R. Jung, and B. J. Betts. Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nature Medicine*, 6(11):1290–1292, Nov 2000.

• [11] T. Sing and N. Beerenwinkel. Mutagenetic tree Fisher kernel improves prediction of HIV drug resistance from viral genotype. In S. B, P. J., and H. T., eds., *Advances in Neural Information Processing Systems 19*, Vancouver, B.C., Canada, 2007, pp. 1–9. MIT.

• [12] M. Zazzi, E. Aharoni, A. Altmann, F. Bazsó, P. Bidgood, G. Borgulya, J. Denholm-Prince, M. Fielder, R. Kaiser, C. Kent, T. Lengauer, T. Nepusz, H. Neuvirth, Y. Peres, A. Petroczi, M. Prosperi, L. Romano, M. Rosen-Zvi, E. Schülter, T. Sing, A. Sonnerborg, R. Thompson, G. Ulivi, L. Zalány, and F. Incardona. Euresist: exploration of multiple modeling techniques for prediction of response to treatment. In *Proceedings of the 5th European HIV Drug Resistance Workshop*, Paris, France, 2007, HIV Medicine. European AIDS Clinical Society.

14.4.2 Analysis of Coreceptor Usage

Investigators: Oliver Sander and Tobias Sing

HIV cell entry and coreceptor usage

HIV virions enter human host cells through consecutive interaction with the CD4 cell surface receptor and one of the two major coreceptors CCR5 and CXCR4. After binding to CD4, a conformational switch in the surface protein gp120 of HIV reveals the coreceptor binding site, most notably, the third hypervariable loop region V3. The V3 loop is considered to be the major viral determinant for coreceptor specificity[3]. After successful attachment to the host cell, fusion of the viral and host cell membranes takes place.

Monitoring coreceptor usage is of great importance due to its relation to disease progression towards AIDS as well as its relevance for therapeutic decisions regarding coreceptor inhibitors. Although phenotypic assays for monitoring coreceptor usage are commercially available, they are time-consuming and costly. To become a routine part of clinical diagnosis, inferring the phenotype from cheaper and faster genotypic analysis is desired.

Established prediction methods like the 11/25 charge rule[4] (predicting CXCR4-usage the in presence of positively charged residues at positions 11 and 25 in the V3 loop) or newer methods based on statistical learning techniques or PSSMs[2] are used to predict coreceptor

tropism based on the V3 loop region of the viral envelope protein gp120, without requiring expensive phenotypic testing.

Previous work on coreceptor usage in the group

The system geno2pheno[coreceptor] is a web service for predicting viral coreceptor usage from the third variable loop of the HIV envelope protein gp120. It is offered along with geno2pheno and geno2pheno[resistance] at http://www.geno2pheno.org. While being exercised by a growing user base since the start in 2004, the tools were also directly applied by us in several collaborations.

The major motivation for recent improvements was to enhance the understanding of coreceptor usage and driving forward the use of prediction-based coreceptor methods in routine clinical practice.

Interpretable predictive models

While support vector machines are widely considered as the state-of-the-art in prediction performance, there is a common attitude that these models are difficult to interpret. Our previously available SVM models were improved by feature ranking.

The weights of individual features in SVM models were obtained by exploiting the bilinearity of the scalar product that defines the linear kernel [6]. Using the linear kernel $k(x, y) = \langle x, y \rangle$ (standard nonlinear kernels did not significantly improve accuracy), feature ranking is particularly straightforward. Due to the bilinearity of the scalar product, the SVM decision function can be written as a linear model,

$$f(x) = \sum_i y_i \alpha_i k(x_i, x) + b = \langle \sum_i y_i \alpha_i x_i, x \rangle + b$$

allowing for direct assessment of the model weights. While the feature rankings are able to reproduce current domain knowledge, they also indicate a prominent role for several novel mutations (see Figure 14.1).

Clinical data from population sequencing and usage of immunological markers

Viral populations *in vivo* are swarms of genetically and phenotypically heterogeneous variants (also termed "quasispecies"). As obtaining a representative number of clonal samples is impractical in clinical practice, the genotype and the phenotype are obtained using bulk or population-based approaches. To address the challenges of population-based data, we developed a strategy for dealing with sequence ambiguities and for integrating clinically derived data into the prediction of coreceptor usage.

Compared to clonal data a substantial performance decrease is observed not only for the 11/25 rule, but also for the SVM with amino acid indicator representation. The inclusion of additional features (CD4% of lymphocytes, number of sequence ambiguities, host CCR5Δ32 heterozygosity, presence of insertions/deletions), possibly acting as surrogate markers for undetected viral subpopulations, leads to considerable improvements in predictive performance. As shown in Fig. 14.2, the improvements in sensitivity over the 11/25 rule and the purely sequence-based SVM are substantial when the clinical parameters are considered.

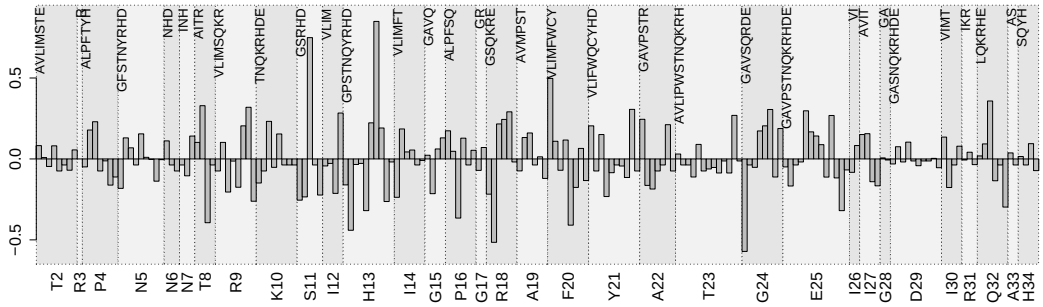


Figure 14.1: Feature ranking. The feature set consists of 700 variables: for each of the 35 V3 positions, 20 variables indicate the presence of specific amino acids. The labels on the horizontal axis indicate the V3 position and the most prevalent amino acid at each position. Bar height represents the weight of a mutation. Only mutations with non-zero weight are shown.

The sensitivity of 63% at the "11/25" specificity corresponds to a 2.4-fold improvement in detecting X4-capable samples relative to the "11/25" rule. The top 5-ranking variables were CD4 percent, presence of 13Y, presence of 11R, number of ambiguous V3 positions, and presence of 24G.

Structural Aspects of Coreceptor Usage

While sequence-based prediction of HIV-1 coreceptor usage showed considerable improvements over the last years, the structural foundations of coreceptor selectivity are still unknown. Recently, a crystal structure of gp120 with its V3 loop was resolved by Huang et al. We use this V3 loop structure as a template to model the V3 loop of viral variants. While the backbone conformation is kept rigid, the SCWRL method[1] is used to predict side-chain conformations. These models of viral variants are then represented by a structural descriptor, using pairwise distance distributions between functional atoms in the loop. This vectorial representation captures the spatial arrangement of physico-chemical properties in the V3 loop and allows to apply statistical learning methods like SVMs and random forests to discriminate between CCR5- and CXCR4-using variants [5].

In 10 replicates of 10-fold cross validation we could show, that the structural descriptor significantly improves prediction of coreceptor usage compared to a linear support vector machine trained on sequence data. For a given specificity of 0.95 a sensitivity of 0.77 was achieved, improving further to 0.80 when combined with a sequence-based representation using amino acid indicators. This compares favorably to the sensitivity of 0.73 for purely sequence-based prediction (Figure 14.3).

By using statistical importance measures, structural features relevant for coreceptor usage can be mapped onto the structure allowing for visual and quantitative interpretation. Future developments will reach on relaxation of the backbone rigidity and improved modeling of V3

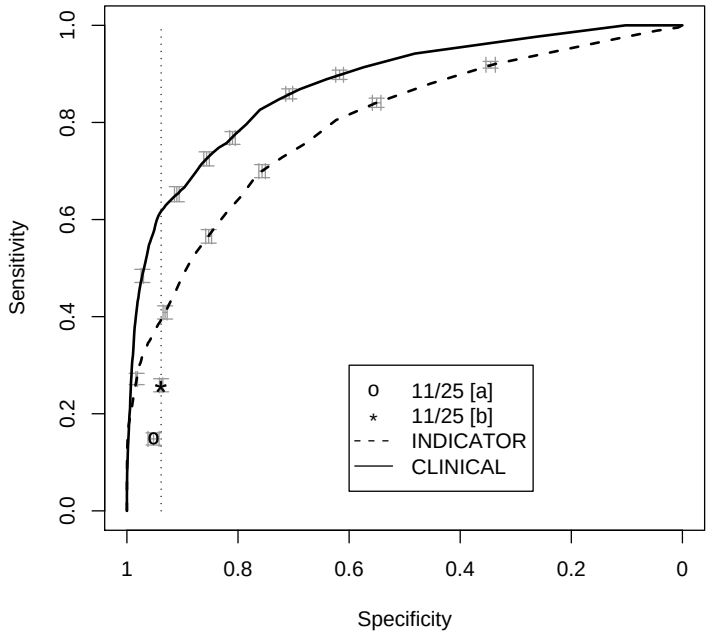


Figure 14.2: ROC curve for prediction of coreceptor usage on population-based data. The dotted vertical line indicates the specificity of the 11/25 charge rule, at which the sensitivities of the methods are compared.

variants with indels.

References

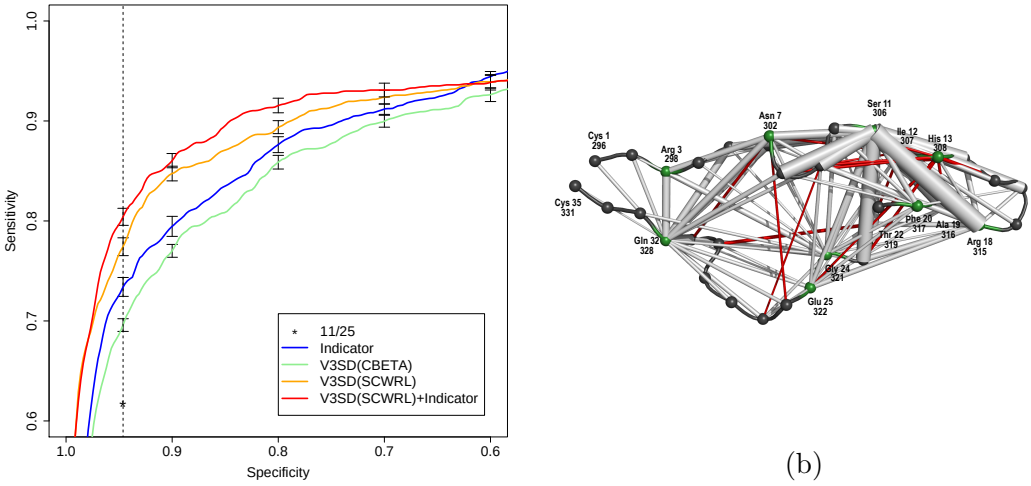
[1] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):2001–2014, 2003.

[2] M. A. Jensen, M. Coetzer, A. B. van 't Wout, L. Morris, and J. I. Mullins. A reliable phenotype predictor for human immunodeficiency virus type 1 subtype c based on envelope v3 sequences. *J Virol*, 80(10):4698–4704, May 2006.

[3] M. A. Jensen and A. B. van 't Wout. Predicting HIV-1 coreceptor usage with sequence analysis. *AIDS Rev*, 5(2):104–112, 2003.

[4] J. J. de Jong, J. Goudsmit, W. Keulen, B. Klaver, W. Krone, M. Tersmette, and A. de Ronde. Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. *J Virol*, 66(2):757–765, 1992.

- [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):e58, 2007.



(a) (b)

Figure 14.3: (a) ROC comparison of predictive performance: sequence-based predictions (*11/25 rule* and *Indicator*) and structural descriptors ($V3SD_{C\beta}$ and $V3SD_{\text{scwrl}}$) on the data set $\text{SEQ}_{\text{noindels},432}$. (b) Importance of residue pairs for donor-aliphatic pseudo-atoms at distances between 9 Å to 18 Å. The width of the edges is proportional to the difference in contributions to the descriptor features of the two coreceptor classes. Edges in gray denote residue pairs characteristic for X4 variants, edges in red mark residue pairs contributing to R5-specific descriptors. Residues which are considered to be important for several distance intervals are marked in green.

- [6] T. Sing, V. Svicher, N. Beerenwinkel, F. Ceccherini-Silberstein, M. Däumer, R. Kaiser, H. Walter, K. Korn, D. Hoffmann, M. Oette, J. K. Rockstroh, G. Fätkenheuer, C.-F. Perno, and T. Lengauer. Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling and SVM-based feature ranking. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, eds., *Knowledge Discovery in Databases: PKDD 2005.*, Porto, Portugal, 2005, pp. 285–296. Springer.

14.4.3 HLA-Virus Interactions

Investigator: Kirsten Roomp

Human immunodeficiency virus (HIV) infection has become a major global human health issue with more than 40 million people infected worldwide and 3.1 million AIDS-related deaths in 2005 alone. A major challenge to natural or vaccine-induced immune control of HIV is the virus' ability to rapidly mutate, when it comes under the pressure of the host's immune system. Antiviral cytotoxic T lymphocytes (CTL) kill HIV infected cells upon recognition of specific viral epitopes. HIV-1 escape mutations interfere with processing of viral antigens by proteasomes or evolve at critical binding sites within the HLA restricted CTL

epitope abrogating binding to the HLA molecule or inhibiting efficient recognition by the T cell receptor. Thus, HIV escapes antiviral immune responses and eradication by the host's immune system.

We have completed a comprehensive analysis of the immune-driven evolution of HIV-1 on an large infected cohort from Bonn. The clinical data of 202 HIV-infected patients, the results of HLA genotyping and sequences of the complete HIV-1 protease and the first half of reverse transcriptase (RT) were analyzed using a variety of statistical approaches. We examined the association of HLA-A, HLA-B and, for the first time, HLA-DRB1 alleles with the emergence of escape mutations in the protease and RT. We also studied the distribution and persistence of escape mutations and their impact on antiviral drug therapy.

We describe new immune escape mutations in both viral protease and RT, some of which are associated with HLA-DRB1, in addition to previously described associations [2]. The escape mutations which were identified are remarkably persistent within our cohort, with a minority of patients developing them more slowly. Interestingly, several HLA-associated escape mutations occur at the same positions as drug resistance mutations. These mutations in our patient cohort were acquired before exposure to these drugs was possible because these drugs were not yet on the market, leading to the conclusion that certain HLA alleles may enhance the development of drug resistance. We examined the distribution of escape mutation *hotspots*, but could not verify them statistically. The influence of HLA on thymidine analogue mutation (TAM) resistance mutation pathways was examined in this cohort using `mtreemix` [1] but no influence of HLA could be found.

Validation of several highly significant escape mutations is not yet possible using existing predication tools. We are planning to use existing structures or homology models, as well as newly available large epitope datasets and statistical learning approaches to help in the verification of these associations. The question of whether HLA alleles that are correlated with a slower disease progression are either more specific (or non-specific) in terms of their binding pocket than those associated with a faster progression will also be looked at. We have identified significant differences in the viral sequences of patients either having Bw4- and Bw6-type alleles which we would also like to verify at a structural level.

References

- [1] N. Beerenwinkel, J. Rahnenführer, R. Kaiser, D. Hoffmann, J. Selbig, and T. Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107, 2005.

[2] C. B. Moore, M. John, I. R. James, F. T. Christiansen, C. S. Witt, and S. A. Mallal. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science*, 296(5572):1439–1443, May 2002.

14.4.4 Recombination Analysis

Investigator: Jochen Maydt

Studying the evolution of pathogens, such as HIV-1, provides many insights into their epidemiology, the evolution of drug resistance and other aspects. However, recombination significantly complicates the analysis of evolutionary processes on the level of molecular sequence

data [1]. It is important to develop methods that infer recombination hotspots, the sequences that are the product of a recombination or whether there is recombination in the alignment at all.

Recombination Analysis Using Cost Optimization

We have developed RECCO [2], a method that analyzes the evolution of sequences subject to mutation and recombination. Given a multiple sequence alignment, RECCO considers all possible reconstructions of one of the sequences from the others by mutation and recombination, where each mutation and recombination involves some cost. RECCO finds all cost-optimal solutions efficiently by dynamic programming. As recombinations can be emulated by mutations and vice versa, RECCO introduces a weight parameter α and recovers all pareto-optimal solutions with a parametric analysis. By comparing the pareto-optimal solution of the lowest and second lowest *alpha* value, RECCO can compute the mutation cost saved by introducing an optimal recombination instead, the so-called *savings* value. The *savings* value is a good indicator for the significance of a recombination signal, but its size depends on the diversity of the alignment. RECCO performs a permutation test based on column permutations of the multiple alignment to obtain an unbiased $p-$value for every recombination event.

In contrast to many recent methods for recombination analysis such as TOPALI [3] or GARD [4], there is no need to choose the size of a sliding window and hence no direct limitation of spatial resolution. RECCO also infers detailed recombination hypotheses for all values of α and presents them in an intuitive graphical way. As RECCO is very fast and intuitive to use, alignments with more than 60 sequences can be analyzed with ease. The recombination detection performance of RECCO was evaluated on synthetic data for several evolutionary scenarios and compared to state-of-the-art approaches. RECCO is superior in recovering recent recombination events, but cannot detect recombinations in duplicated or highly similar sequences, by design. On random genealogies, RECCO performs comparably to state-of-the-art approaches.

Handling Gaps in RECCO

Most recombination analysis methods do not discuss how gaps are incorporated into the algorithm and assume that some appropriate method is used. However, different ways to treat gaps can have a severe impact on the output of the recombination analysis program and lead to the inference of spurious recombination events.

We have evaluated potential gap cost functions and identified several restrictions that a gap cost function for RECCO must satisfy. For example, the permutation test requires a gap cost function without fixed cost terms. Based on a mathematical analysis of the cost function in RECCO, it is also possible to derive a lower and an upper bound for the cost of a gap. Scoring gaps outside these bounds results in optimal solutions that are biologically not relevant. For example, the optimal solution may include or exclude the gap irrespective of the sequence surrounding the gap. It is particularly promising to score gaps with a cost close to the lower bound. In this case, a gap is only included into the optimal solution if the sequence surrounding the gap fits the putative recombinant at least as well as some other sequence. However, it is difficult to maintain the interpretation of the *savings* value in the

parametric analysis.

In contrast to other common approaches for handling gaps, it is not necessary to discard sequence information. There is also a lower risk of assigning a gap cost that forces a particular solution and introduces spurious recombination events.

References

[1] P. Awadalla. The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics*, 4(1):50–60, Jan 2003.

- [2] J. Maydt and T. Lengauer. Recco: recombination analysis using cost optimization. *Bioinformatics*, 22(9):1064–1071, February 2006.

[3] I. Milne, F. Wright, G. Rowe, D. F. Marshall, D. Husmeier, and G. McGuire. TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments. *Bioinformatics*, 20(11):1806–7, Jul 2004.

[4] S. L. K. Pond, D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. W. Frost. GARD: a genetic algorithm for recombination detection. *Bioinformatics*, Nov 2006.

14.5 Protein Structure and Function Prediction

Proteins perform key roles in all living systems. They can be characterized at sequence, structure, and functional level. This section of the report covers structural and functional aspects, such as structural characterization, analysis of structure-function relationships, and investigation of functional relationships.

Structural characterization is the focus of two projects. Assessing the quality of predicted protein structures is described in Subsection 14.5.1, and the structural characterization of alternative experimental models is described in Subsection 14.5.2. We analyze structure-function relationships, focusing on protein-protein interactions. In this respect, we currently develop methods to compare protein-protein binding sites 14.5.3 and interfaces 14.5.4. Functional relationships are investigated in two additional projects. Statistical learning methods based on structural and sequence similarity are applied in one of these projects to infer and predict functional annotations 14.5.5. In the other project, functional relationships are analyzed based on metabolite profiling in yeast 14.5.6.

These projects are complemented by the work described other sections. We actively cooperate with the molecular networks team 14.6 as we share common interests regarding protein interactions. We closely collaborate in methods development with the computational chemical biology team 14.8. We continuously cooperate with other teams regarding medical applications. To illustrative examples are the work on HIV 14.4.2 and HCV 14.6.3.

14.5.1 Improving Model Quality in Structure Prediction

Investigators: Ingolf Sommer and Oliver Sander

In the area of protein structure prediction, recently there has been considerable activity in the development of Model Quality Assessment Programs (MQAPs). MQAPs distinguish high quality protein structure models from inferior models. We proposed a new method to use an

MQAP to improve the quality of models[1]. In template-based protein structure prediction alignment is known to be a bottleneck limiting the overall model quality. With a given target sequence and template structure, a number of different alignments and corresponding models were constructed for the sequence. The quality of these models was scored with an MQAP and used to choose the most promising model. An SVM-based selection scheme was suggested for combining MQAP partial potentials, in order to optimize for improved model selection.

The approach has been tested on a representative set of proteins. The ability of the method to improve models was validated by comparing the MQAP-selected structures to the native structures with the model quality evaluation program *TM*-score. Using the SVM-based model selection, a significant increase in model quality was obtained (as shown with a Wilcoxon signed rank test yielding p-values below 10^{-15}). The average increase in *TM*score is 0.016, the maximum observed increase in *TM*-score is 0.29. We showed that a combination of systematic alignment variation and modern model scoring functions can significantly improve the quality of alignment-based models.

References

- [1] I. Sommer, S. Toppo, O. Sander, T. Lengauer, and S. Tosatto. Improving the quality of protein structure models by selecting from alignment alternatives. *BMC Bioinformatics*, 7:1–11, July 2006.

14.5.2 Analysis of Structural Variants of Proteins

Investigators: Francisco Domingues and Jörg Rahnenführer

The increasing availability of structural information provides a unique opportunity for better understanding and predicting protein function at the molecular level. Structural models for different types of proteins are constantly being added to public databases where they are stored, annotated and classified in a systematic way. At the same time, alternative structural models for the same protein are also becoming available. These alternative models can display considerable structural differences. A systematic analysis of the differences between these alternative models is essential for modelling and structure prediction, for ligand and protein docking, for conformational analysis and for investigating structure-function relationships. We have developed STRuster, a method for the analysis and characterization of alternative protein structures.

STRuster originally integrated a clustering procedure for grouping alternative models according to backbone structure similarity [1]. The method has now been extended to identify the flexible and the invariant regions in the protein backbone, and to characterize the structural variability in functional sites.

Plots of the standard deviation of atomic distances between backbone atoms are used to visualize the structurally conserved (invariant) regions, the hinge segments associated with local structural changes, and the structurally variable segments. A data smoothing approach is applied in order to identify these different regions. In addition, subsets of alternative models are compared in order to identify and locate distinct conformational states. The method takes into account estimates of atom coordinate uncertainty computed for most of the

structures determined by X-ray crystallography. The invariant regions are used to define sets of equivalent residues between the different models and to generate optimal superpositions, see Figure 14.4. STRuster was also extended to characterize sites of interest in the protein, like ligand binding sites, catalytic sites or protein protein interaction sites. In particular the coordinates of the side chain atoms are used to cluster the models according to the structural similarity and to assess the degree of structural variability at these sites.

The method was applied to the models determined by x-ray crystallography, and considerable structural variability was observed for most proteins. The results are available on the STRuster web site `http://struster.bioinf.mpi-inf.mpg.de/`. STRuster is currently being applied to investigate the structural basis for viral drug resistance. In particular we are analyzing how mutations in protease NS3 from hepatitis C virus (HCV) confer resistance to a VX-950, a known protease inhibitor with strong antiviral activity [2], see also Section 14.6.3.

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, June 2004.

[2] C. Lin, K. Lin, Y.-P. Luong, B. G. Rao, Y.-Y. Wei, D. L. Brennan, J. R. Fulghum, H.-M. Hsiao, S. Ma, J. P. Maxwell, K. M. Cottrell, R. B. Perni, C. A. Gates, and A. D. Kwong. In vitro resistance studies of hepatitis c virus serine protease inhibitors, vx-950 and biln 2061: structural analysis indicates different resistance mechanisms. *J Biol Chem*, 279(17):17508–17514, Apr 2004.

14.5.3 Analysis of Protein Binding-sites

Investigators: Oliver Sander, Francisco Domingues, and Ingolf Sommer

Motivation

Protein-protein interactions are the basis of many cellular processes. While a considerable number of such interactions are known, molecular details are missing for many of them. As experimental high-throughput methods for determining binary interactions produce a significant number of false positives, molecular and structural details are desired to confirm and validate networks of interacting proteins. Furthermore, as structural aspects are the determinants of recognition, specificity, and affinity, descriptive and predictive methods are needed to cope with the scarcity of newly determined structural complexes[3]. See Figure 14.5 for an exemplary analysis of the structural determinants of specificity in G-protein binding sites.

In a recent study (see Section 14.4.2) we applied structural descriptors to the problem of HIV-1 coreceptor usage. In this study descriptors of a binding site were used to predict the binding partner. The generalization to a variety of binding sites could be used for surface comparison, prediction of similar binding partners, analysis of mimicking, and protein function prediction. Using a vectorial representation the binding site descriptions are tractable by a wide variety of statistical learning and multivariate analysis tools.

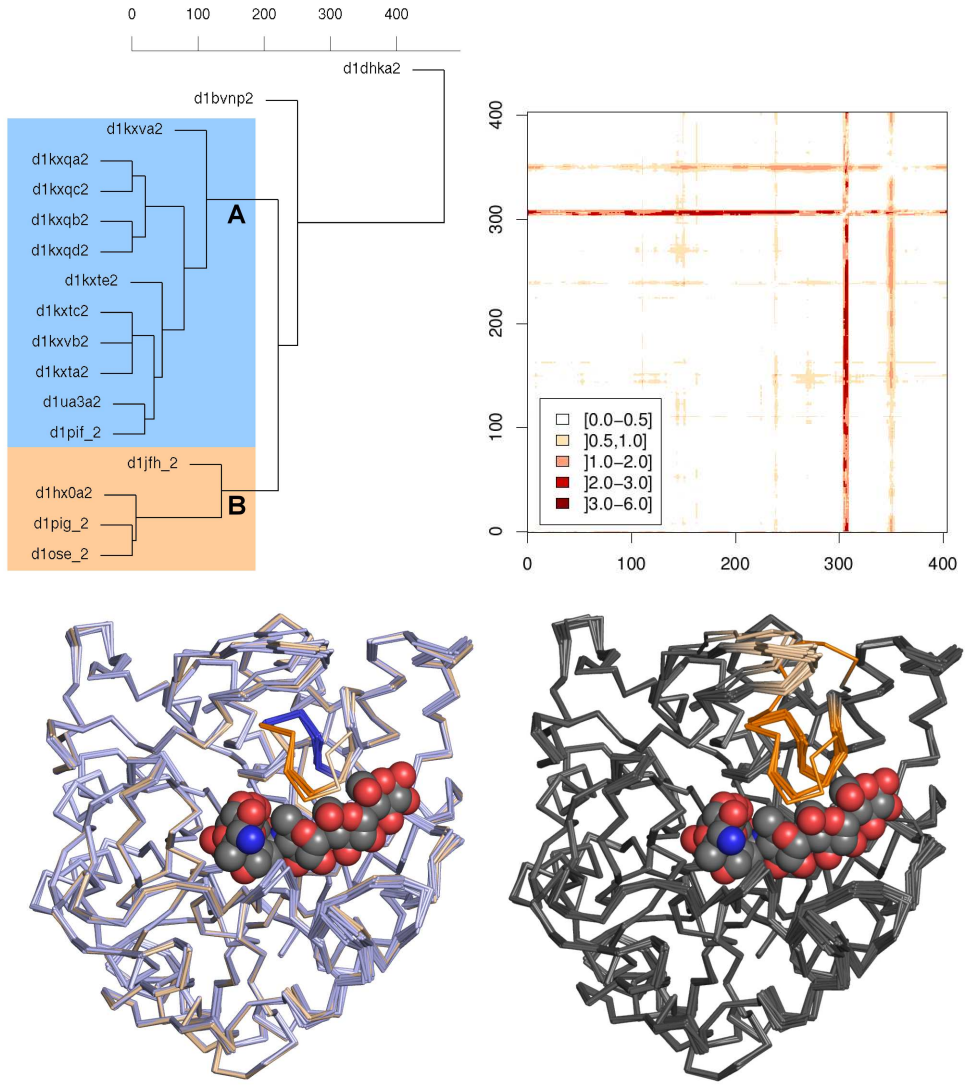
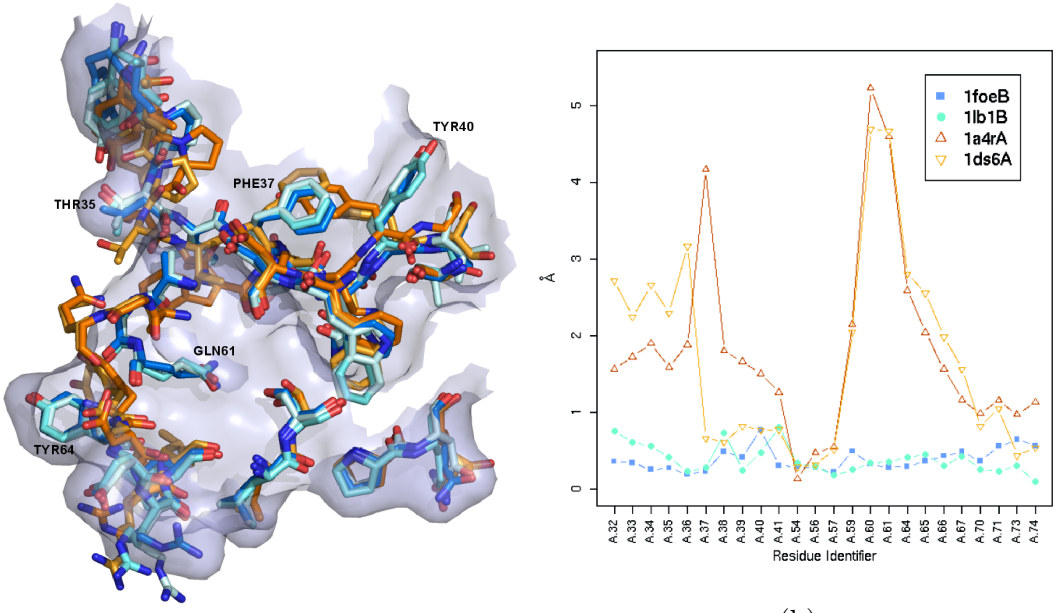


Figure 14.4: Analysis of 17 structural models of α-amylase. Two main clusters (A and B) are observed in the dendrogram obtained with hierarchical clustering. Comparison between models from cluster A to models from cluster B using STRuster reveals that the loop around position 305 adopts two different conformations. The two alternative conformations of this loop are visible on the bottom left superposition of the different models. The loop (blue) is in an open conformation in cluster A where the active site is empty, but is closed (orange) in the models from cluster B where a ligand is filling the active site. The matrix of the standard deviation of the C^{α} atom distances (top right) shows two loops with high structural variability around positions 305 and 350. These loops surround the ligand binding site, and are shown in orange on the superposition on the bottom right.



(a) (b)

Figure 14.5: Comparison of five binding sites of the G-protein. (a) The surface of the reference binding site (1kilA) is in blue and the interacting residues are represented as sticks. This binding site was compared to four other binding sites regarding the RMSD after optimal superposition of the Cα atoms. (b) Cα atom distances between the reference binding site from 1ki1A and the other four binding sites after superposition. The three binding sites which interact with the a.87.1.1 protein family (1ki1A, 1foeB and 1lb1B, in blue) have the highest structural similarity. The two other binding sites, 1a4rA and 1ds6A (in orange), show larger structural differences, in particular around residues 37 and 64.

Methods

Defining Binding Sites. Binding site residues are determined by the ΔASA criterion defined by Jones and Thornton [1]. The residues can be represented by their Cα or Cβ atoms or by more sophisticated representations, for example representing the side chains of residues at a protein binding site by five functional atom types (see Figure 14.6). Following [4], these atom types are labelled as hydrogen-bond donor, acceptor, ambivalent donor/acceptor, aliphatic, or aromatic ring.

Describing Binding Sites. Currently, we are investigating two variants of structural descriptors for representing the geometric arrangement of important atoms or functional groups: distance distributions and three-dimensional moments[2].

To compute the distance distributions the distances between all pairs of functional atoms are computed and aggregated into a probability density for each type pair. Thus, when using 5 functional atom types each binding site is represented by a set of 15 distribu-

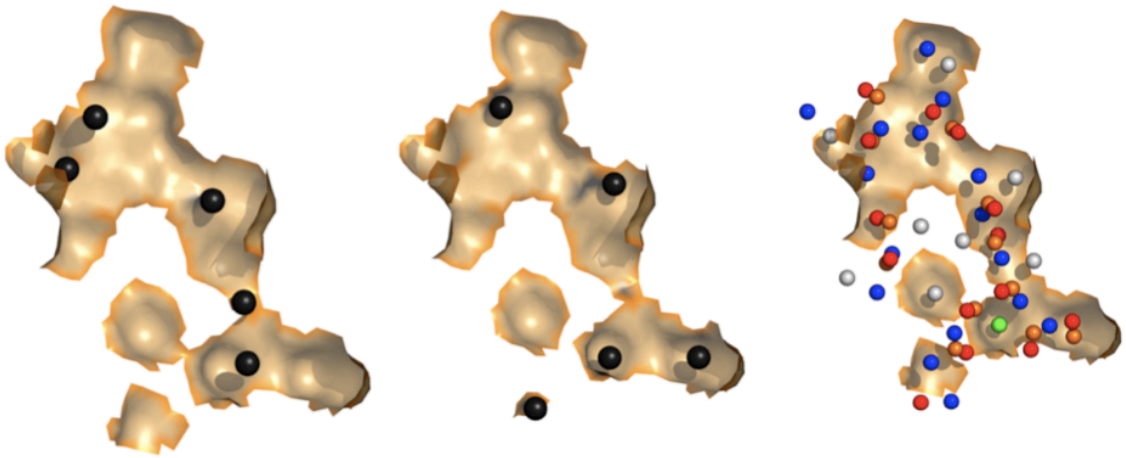


Figure 14.6: Protein-protein binding site of 1cdt represented by Cα-atoms (left), Cβ-atoms (middle), and functional atoms (right). The functional atoms according to [4] are colored blue (donor), red (acceptor), green (ambivalent donor/acceptor), white (aliphatic), and orange (pi-stacking). The solvent accessible surfaces of the binding site residues are visualized in amber.

tions. For comparison of the binding site descriptors L_1 and L_2 norms, cosine distance, Kolmogorov-Smirnov distances, or information theoretic measures like Kullback-Leibler or Jensen-Shannon divergence can be used.

Moments are used to describe the geometric arrangement of possibly labelled points in three-dimensional space. From these moments features can be extracted that are invariant under translation and rotation and are thus suited for vectorial description of binding site residues.

While the moment-based descriptors provide a very compact representation, distance distributions allow for an intuitive handling of binding site flexibility by thresholding the maximal distances to be considered.

Results and Future Directions

Current versions of the structural descriptors allow for retrieving similar binding sites from a large set of sites with an AUC of above 80% on average (see Figure 14.7). Considering that assigning binding sites to groups of known high similarity is a multi-class classification problem with dozens or even hundreds of classes the performance seems adequate and suited for application. As both descriptors currently describe the global geometric arrangement of the binding site, changes in the set of binding residues lead to considerable changes in the descriptors. Currently a proof of concept is being developed using structural descriptors to represent local parts of protein surface independent from small changes in binding site boundaries.

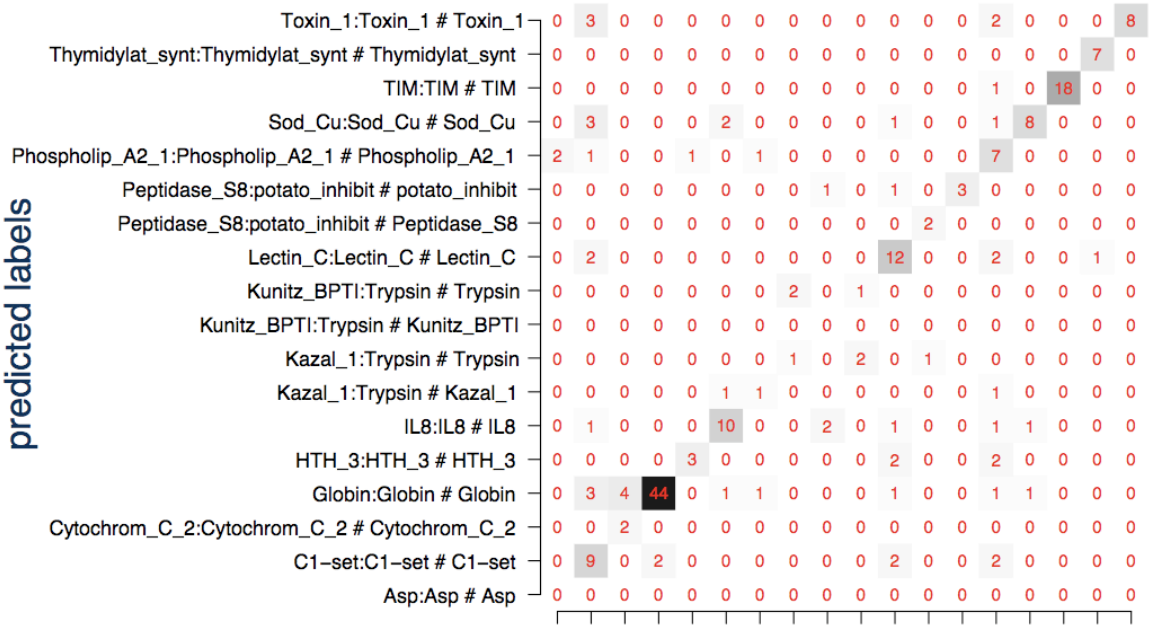


Figure 14.7: Confusion matrix for a set of diverse binding site groups (data set based on [1]). Entries on the diagonal are correct predictions, whereas off-diagonal entries represent binding sites assigned to wrong families. While the larger classes mostly show good performance, some of the smaller classes are mispredicted.

References

[1] S. Jones and J. Thornton. Protein-protein interactions: a review of protein dimer structures. *Prog Biophys Mol Biol*, 63(1):31–65, 1995.

- [2] O. Mueller. Using shape retrieval techniques for identifying similar protein binding sites. Masters thesis, Universität des Saarlandes, September 2007.

[3] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, 14(3):313–24, Jun 2004.

[4] 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.

14.5.4 Analysis of Protein Interfaces

Investigators: Hongbo Zh, Ingolf Sommer, and Francisco Domingues

Proteins interact with each other through their binding sites. Whereas the investigation and comparison of individual binding sites has been described in the previous section (see 14.5.3), here we focus on two interacting binding sites, forming an interface. We have implemented an approach for characterizing and classifying different protein-protein interaction types based

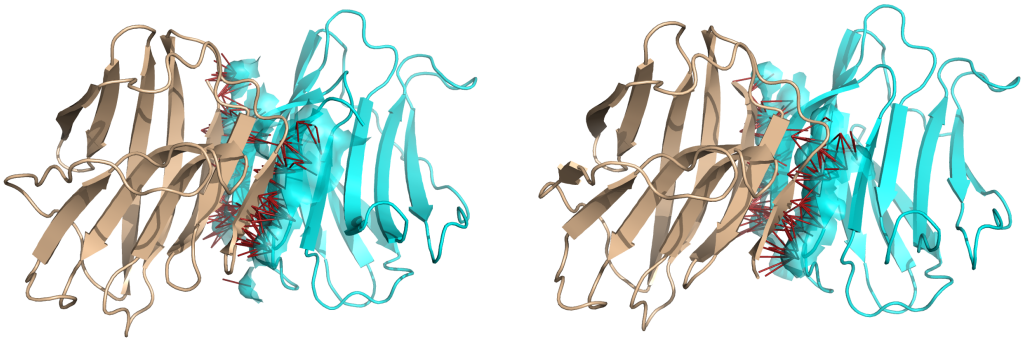


Figure 14.8: Comparison of similar protein-protein interfaces. Non-covalent interactions between binding sites (shown as red lines) are aligned to infer the similarity between the interfaces. Human galectin-1 (PDB code: 1gzw) is on the left; bovine galectin-1 (PDB code: 1slt) is on the right.

on several interface properties. Now, we are developing methods for comparing protein-protein interfaces.

There are different types of biological interactions [2] as well as non-biological interactions presented in protein structural models. Protomers from obligate complexes do not exist as stable structures *in vivo*, whereas protomers of non-obligate complexes may dissociate from each other and stay as stable independent units. Obligate and non-obligate proteins have been shown to have distinct interaction preferences [1]. In order to characterize and differentiate these protein complexes, we have developed *NOXclass*, a classifier identifying protein-protein interaction types (biological obligate, biological non-obligate and crystal packing) implemented using a support vector machine (SVM) algorithm [3]. We have investigated several interface properties for a set of non-redundant protein-protein interactions and found that the most discriminative properties are interface area, interface area ratio, and area-based amino acid composition of the interface. An SVM classifier has been trained with these interface properties to differentiate not only biological interaction from crystal packing contacts, but also obligate interactions from non-obligate interactions. Our SVM classifier achieved an accuracy of 91.8% using leave-one-out cross-validation on a non-redundant protein-protein interaction dataset. A web server based on the *NOXclass* method is made available [1]. Users can employ it to predict the types of given protein-protein interactions. The *NOXclass* implementation can be downloaded from this web server.

We are currently developing a graph-based method to compare protein interfaces. This method takes geometric and chemical properties into account in order to align and measure the similarity between interfaces (see Figure 14.8).

[1] `http://noxclass.bioinf.mpi-inf.mpg.de/`

References

[1] R. P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, 336(4):943–55, Feb 2004.

[2] I. M. A. Nooren and J. M. Thornton. Diversity of protein-protein interactions. *EMBO J*, 22(14):3486–92, Jul 2003.

• [3] H. Zhu, F. S. Domingues, I. Sommer, and T. Lengauer. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7:1–15, June 2006.

14.5.5 Function Prediction Based on Conserved Regions

Investigators: Francisco Domingues, Oliver Sander, and Ingolf Sommer

There are many protein sequences and structures lacking functional annotations. Experimental function elucidation is a tedious task which can be facilitated using function predictions. The majority of current computational methods employs global and local sequence and structure (plus combinations of these) to infer protein function.

We developed a method for predicting protein function, based on a concept called 'functionally conserved regions' [2]. Our strategy identifies local regions in sequence and structure space which are composed of proteins sharing the same functions. For each Gene Ontology (GO) term such regions of conserved function are identified. These regions are stored along with the information how likely a certain function is attained in neighborhoods around the regions. For functionally unlabeled proteins this information can be used to predict the relevant GO terms. Contrary to previously existing function prediction methods, we use several nearest neighbors; see Figure 14.9 for an overview of the method.

In order to assess the performance of this function predictor, we developed an evaluation which can be applied to any method that produces a ranking of GO terms. The procedure neither requires restrictions to specific levels of the GO DAG to measure the degree of correctness. Therefore, it is applicable to any prediction method independent of the strategy used.

The use of 'functionally conserved regions' is an extension of the traditional view on sequence and structure space. We compare our method to state-of-the-art function predictors. For example, when predicting functions of GO level two we achieve an accuracy of 92% per protein compared to 85%-90% in [1].

References

[1] 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, Oct 2004.

• [2] N. Weinhold. Inference of protein function based on functionally conserved regions in sequence and structure space. Masters thesis, Universität des Saarlandes, December 2006.

14.5.6 Analyzing Metabolic Networks in Yeast

Investigator: Priti Talwar

The yeast, *Saccharomyces cerevisiae*, is a simple eukaryotic organism with approximately 6000 genes. *Saccharomyces cerevisiae* is an ideal model organism for large-scale functional

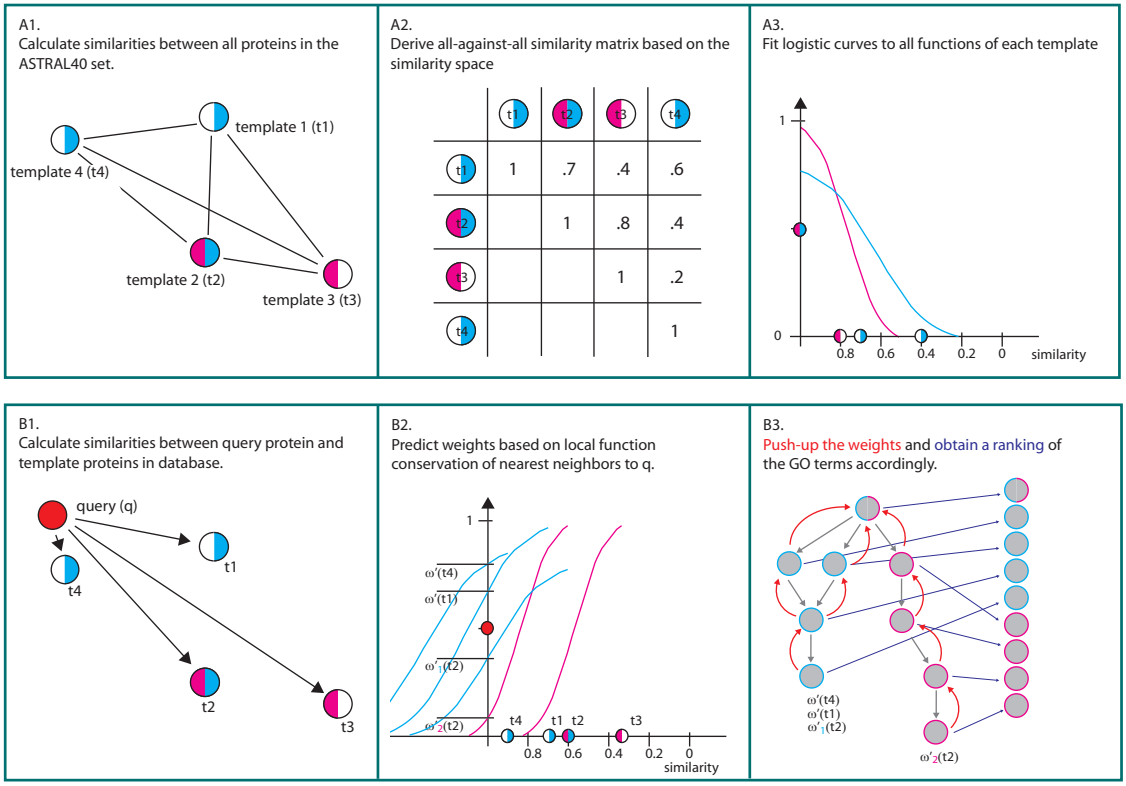


Figure 14.9: Overview of the prediction procedure. The blue and pink colors represent two distinct functions. Upper figure: preprocessing steps. The method uses several sequence or structure similarity spaces. Here, the idea is exemplified using only one such similarity space. In this space, we calculate similarities for all proteins in the dataset (A1), which gives us all pairwise similarities (A2), we fit logistic curves for all GO functions and templates (A3)(only one template with two GO functions is shown). Lower figure: the actual prediction of GO function. For a query protein q, the similarities to all templates in the dataset are calculated (B1). These similarities to the nearest neighbors are used to look up the weights from the neighbors fits (those we fitted in A3) (B2). Finally, we map the functions, with the weights to the GO graph, and obtain a ranking of the functions (B3).

studies and provides a system in which genes can be systematically inactivated by way of gene-knockout methods [2]. A substantial fraction of the 6000 genes in *Saccharomyces cerevisiae* encode proteins for which currently we do not know any confirmed or putative function. Prediction of the functional role of these proteins is a challenging problem in systems biology, especially when many of these genes have no overt phenotypes. Here we aim at a better understanding of the underlying functional relationships between genes working across diverse metabolic pathways using intracellular metabolite profiling studies.

We are applying bioinformatics methods and statistical analysis techniques in combination with stoichiometric and flux profiling to understand the function and the regulatory mechanisms of specific genes involved in central carbon metabolism and amino acid biosynthesis, see Figure 14.5.6. The experimental work is carried out by the group of Prof. Elmar Heinzle (Biochemical Engineering, Saarland University), our collaboration partner. ^{13}C stable isotope substrates can be used as tracers to generate detailed metabolic profiles of gene knockouts. Detailed and quantitative information on the physiological cellular states is measured by ^{13}C-metabolic flux analysis of cultures grown in high throughput novel oxygen sensor microtiter plates. In this project, we aim to develop a systematic approach for study of *Saccharomyces cerevisiae* genes of unknown function based on the metabolic profiles of knockout mutants grown under different carbon substrates. As a first step, we have developed a tool for automation of Gas Chromatography Mass Spectrometry data acquisition and analysis routine as this is a bottleneck in the metabolic profiling studies [8]. According to our hypothesis, similarity in the metabolic profiles can be used to find functionally linked genes. *Saccharomyces cerevisiae* is known to be robust to majority of genetic perturbations. In the next step, we worked on finding closely related mutants which show similar metabolic profiles (stoichiometric and labeling profiles). In these cases where the mutants show no overt phenotypes, we developed a sensitive method to detect those subsets of metabolic profile features which are most differentiating (outliers) for all mutants. Furthermore, in the recent year's genome scale metabolic models are being developed. We have built a web-server which integrates metabolic profiling data and performs topological analysis on custom-defined and genome scale models of *Saccharomyces cerevisiae*.

Gas Chromatography Mass Spectrometry Analysis

In the group of our collaboration partners, ^{13}C stable isotope substrates are used during the cultivation to trace the intracellular molecular fragments using Gas Chromatography Mass Spectrometric (GC-MS) studies [4, 6, 9]. For each mutant, we use data on sfl and stoichiometric profiles. The sfl profile is calculated from mass isotopomer labeling of various amino acids and their corresponding fragments. The stoichiometric profile for a mutant is represented by six features in our analysis namely growth rate (/h) X, biomass yield (g/ g glucose) Y_{xs}, ethanol yield (g/g) Y_{ps}, biomass yield on O2(rich)culture condition (g/g) Y_{xo}, rate of production of biomass ((g/g)/h) q_s, rate of production of ethanol ((g/g)/h) q_p, see Figure 14.5.6.

Classification of mutants based on metabolic profiles

For a specific set of mutants, we applied partition around Medoids (PAM) and hierarchical clustering (HC) methods. The silhouette width was calculated and used as a measure of cluster quality. M_{glu} is the set of 37 mutants which are grown under glucose conditions; M_{fru} is the set of 41 mutants which are grown under fructose conditions and M_{gal} is the set of 24 mutants which are grown under galactose conditions. The stoichiometric profiles under glucose and fructose conditions show high similarity. M_{glu}, M_{fru} and M_{gal} lead to silhouette width values of 0.31, 0.36, and 0.33 using PAM, and, 0.36, 0.31, 0.42 using HC, respectively. Stoichiometric profiles provide global features which must be complemented with other heterogeneous data types, for mutant differentiation. The above set of mutants was also evaluated for association at the level of co-response data [5]. We found that 33% of the highly correlated co-response profile mutant pairs also belong to the same cluster as given by stoichiometric profile analysis using PAM and HC algorithm. Out of 630 total pairs, 185 mutant pairs show correlation greater than equal to 0.5 on the co-response profile level. Out of these, 29 mutant pairs show co-response profile correlation of greater than or equal to 0.7. Preliminary investigation of the metabolic profile data and the co-response data showed that the combination of these heterogeneous data renders better resolution of the functional associations among mutants, but does not lead to well separated clusters. Currently, we are working on outlier detection method for identifying the most differentiating features of a mutant as well as the outlier mutants in a given mutant set, [7] see Figure 14.5.6.

The relationship between the topology and function of metabolic networks is of central interest in computational systems biology. Therefore, we have developed a web server for the analysis of metabolic profiling data in genome scale *Saccharomyces cerevisiae* metabolic models, as well as user-defined models [1, 3].

References

[1] N. Duarte, M. Herrgard, and B. Palsson. Reconstruction and validation of saccharomyces cerevisiae ind750, a fully compartmentalized genome-scale metabolic model. *Genome Res.*, 14(7):1298–309, 2004.

[2] K. Entian, T. Schuster, J. Hegemann, D. Becher, H. Feldmann, U. Guldener, R. Gotz, M. Hansen, C. Hollenberg, G. Jansen, W. Kramer, S. Klein, P. Kotter, J. Kricke, H. Launhardt, G. Mannhaupt, A. Maierl, P. Meyer, W. Mewes, T. Munder, R. Niedenthal, R. M. Ramezani, A. Rohmer, A. Romer, and A. Hinnen. Functional analysis of 150 deletion mutants in saccharomyces cerevisiae by a systematic approach. *Mol. Gen. Genet.*, 262:683–702, 1999.

[3] J. Forster, I. Famili, P. Fu, B. Palsson, and J. Nielsen. Genome-scale reconstruction of the saccharomyces cerevisiae metabolic network. *Genome Res.*, 13(2):244–53, 2003.

[4] P. Kiefer, E. Heinzle, O. Zelder, and C. Wittmann. Comparative metabolic flux analysis of lysine producing corynebacterium glutamicum cultured on glucose or fructose. *Appl. Environ. Microbiol.*, 70(1):229–239, 2004.

[5] D. Steinhauser, B. Usadel, A. Luedemann, O. Thimm, and J. Kopka. Csb.db: a comprehensive systems-biology database. *Bioinformatics*, 20(18):3647:51, 2004.

[6] P. Talwar, T. Lengauer, C. Wittmann, V. Mangadu, and E. Heinzle. Towards cellular function through metabolite screening. In *Proceedings of the 3rd Annual Conference on Metabolic Profiling: Pathways in Discovery*, Princeton, New Jersey, 2003, p. 7. Cambridge Healthtech Institute.

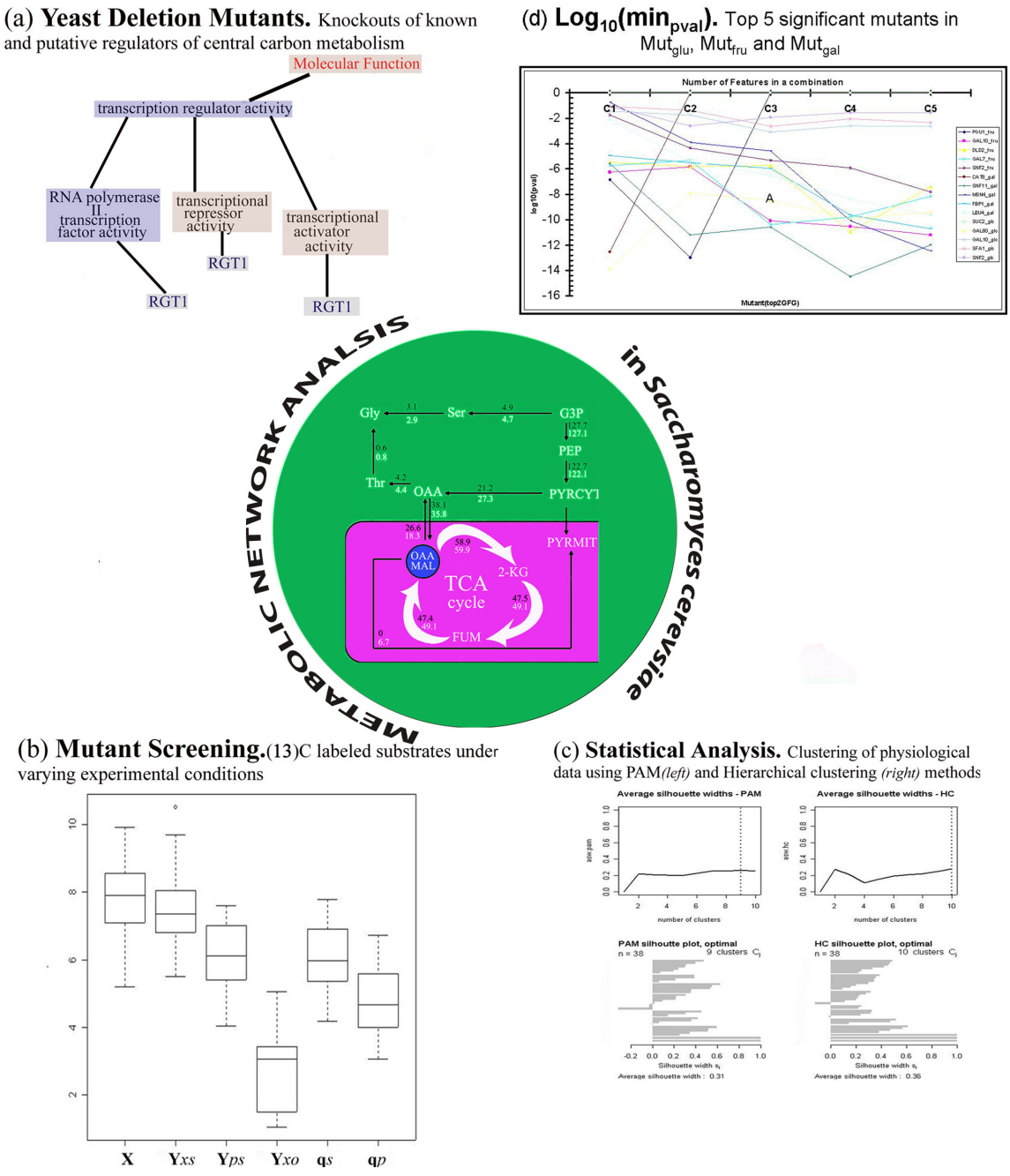


Figure 14.10: Workflow for computational analysis of metabolomics data for protein function prediction in *Saccharomyces cerevisiae*. (*a*) GO function terms associated with the RGT1 gene product. (*b*) Distribution of physiological yield data of the six considered features. (*c*) Results of clustering the mutants. The graphs on the top show the average silhouette-widths in dependence on the number of clusters. The histograms on the bottom show the silhouette widths of knockouts in an optimal clustering. (*d*) $\log_{10} \min_{pval}$ for top five significant mutants in each mutant set namely M_{glu}, M_{fru} and M_{gal} in dependence on the number of features in a feature combination. $C1$, $C2$, $C3$, $C4$ and $C5$ are feature combinations with 1, 2, 3, 4 and 5 features, respectively. M_{glu},M_{fru} and M_{gal} are the set of 37, 41 and 24 mutants which are grown with glucose, fructose and galactose substrates, respectively. min_{pval} is the p-value which is minimum of the p-values for the combinations with 1, 2, 3, 4 and 5 features.

- [7] P. Talwar, T. Lengauer, C. Wittmann, V. Velagapudi, and E. Heinzle. Development of computational methods for analysis of metabolic profiling data. In *Proceedings of International Conference on Systems Biology (ICSB 2006), Japan*, Yokohama, Japan, 2006, pp. 42–42. International Conference on Systems Biology 2006.

[8] P. Talwar, C. Wittmann, T. Lengauer, and E. Heinzle. Software tool for automated processing of 13C labeling data from mass spectrometry data. *BioTechniques*, 35:1214–1215, December 2003.

- [9] V. Velagapudi, C. Wittmann, T. Lengauer, P. Talwar, and E. Heinzle. Metabolic high- content screening of saccharomyces cerevisiae reveals superior performance of single gene deletion strains. *Process Biochemistry*, 41:2170–2179, 2006.

14.6 Molecular Networks in Medical Bioinformatics

Coordinator: Mario Albrecht

The new research group Molecular Networks in Medical Bioinformatics concentrates on the study of molecular interaction networks and the prediction of protein function. Recently, our research focused on the analysis of the human interactome and experimentally derived and predicted protein-protein interactions. However, previous work has also been continued regarding the development and application of various bioinformatics methods to extract knowledge about the function and structure of proteins with medical relevance. To this end, the research group has been cooperating with about ten biological and medical groups from Europe and USA.

The research group headed by Mario Albrecht (formally established on 1 November 2006) currently comprises two PhD students and one post-doc who collaborate on the projects detailed in the following three subsections. During the last two years, six master students have worked on their master theses in the group. After earning their master degree, two of the students will remain in the group to continue their work on the impact of alternative splicing on protein networks and on an Internet exchange protocol for molecular interaction data.

The projects conducted in this research group are closely related to work on protein structure and function described in the Section 14.5. This includes active cooperation regarding the use of molecular networks in the study of structural and functional aspects of proteins. In addition, our research on hepatitis C is supported by the application protein-ligand docking 14.8 in joint case studies.

Subsection 14.6.1 describes joint work with Section 14.5 on the development of a novel functional similarity measure based on the Gene Ontology (GO) and its application to protein-protein and domain-domain interactions. This work was presented at the German Conference of Bioinformatics in 2006. It also includes the implementation of the GOTax software platform that can be used by biologists for investigating the distribution of specific protein families or functional GO terms over the taxonomic tree.

Subsection 14.6.2 details the comprehensive evaluation of publicly available datasets of human protein-protein interactions, which was presented at the Interactome Networks Meeting in 2006. The analysis results provided important insights into the qualitative differences between experimental and predicted data. In particular, the findings enable the derivation of appropriate confidence scores for the selection of reliable interactions, which will form

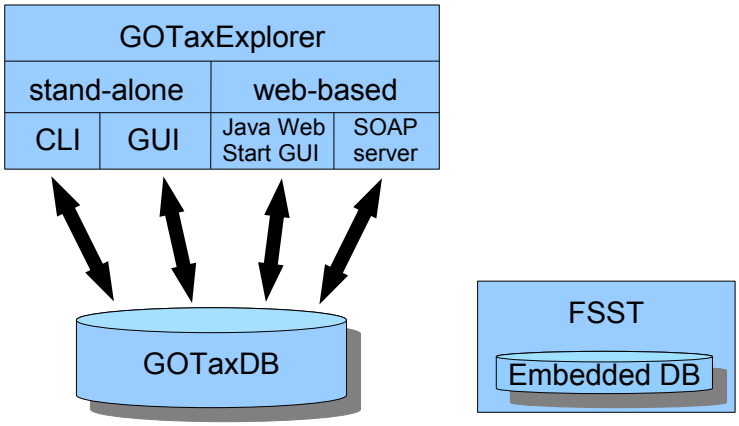


Figure 14.11: Schematic drawing of the GOTax platform. The Web Start version of GOTaxExplorer does not require the user to install a local copy of GOTaxDB, but uses an online version instead. The SOAP server provides the possibility of integrating the GOTax platform into other services without the need of installing a local copy.

the basis of further disease-related network studies. Apart from that, the NetworkAnalyzer plug-in for Cytoscape was developed for computing topological network parameters and for investigating networks of polyanion-binding proteins.

Subsection 14.6.3 explicates application studies of bioinformatics methods for predicting the structure and function of disease-related proteins. Medical cooperation partners identify promising candidate proteins that cause specific diseases, but are as yet relatively uncharacterized. The results of our bioinformatics analysis support the prioritization of experiments targeted towards elucidating the molecular function of disease proteins. Currently, we focus on autoinflammatory and neurodegenerative disorders (e.g. Crohn disease and Parkinson) and viral diseases such as hepatitis C caused by flaviviridae.

14.6.1 Functional Similarity of Proteins and Domains

Investigators: Andreas Schlicker, Mario Albrecht, and Francisco S. Domingues

An increasing number of genomes from species across the whole taxonomic tree have been sequenced and extensively annotated. As a result, there is a new opportunity for investigating molecular biology at a taxonomic level, by comparing sets of genomes in a systematic way. Gene Ontology (GO) is widely accepted as standard for functional annotation of proteins and protein families. It can be used to infer functional relationships between proteins and protein families, which complements traditional homology-based methods.

We developed a measure for the functional similarity of gene products, which exploits the GO annotation categories molecular function and biological process [3]. We first defined a measure of semantic similarity between GO terms, sim_{Rel}, that is based on measures from Resnik [2] and Lin [1]. We then extended sim_{Rel} and developed a score of functional similarity between gene products ($funSim$). This similarity score can be used to establish direct functional relationships between proteins from different species independent of homology. The

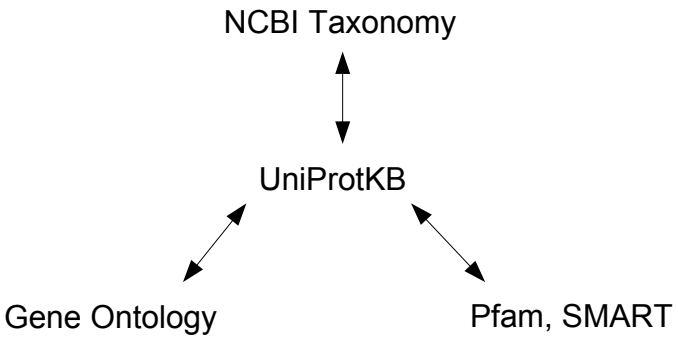


Figure 14.12: Information triangle showing the data included in GOTaxDB.

sim_{Rel} and $funSim$ scores can be applied for comparing the molecular biology of different groups of organisms along the taxonomic tree. They are also useful for revealing functions or processes shared between groups or unique to one group. Another possible application is the identification of proteins from pathogens that are not functionally similar to human proteins, discovering potential new drug targets.

To address these questions, we implemented the GOTax platform for investigating the distribution of specific protein families or functional term over the taxonomic tree [5]. GOTax consists of GOTaxDB, GOTaxExplorer, and FSST (Figure 14.11). GOTaxDB integrates protein annotation from UniProt, protein family classification using Pfam and SMART, functional terms from GO, and the NCBI Taxonomy (Figure 14.12). These different data sources are cross-linked, allowing users to query for all available data. An example is the query "Which biological processes are annotated to proteins from fungi with a PHP domain?". GOTaxExplorer is the query interface to GOTaxDB. It is available for download including a graphical user interface and a command line interface, and as online version (`http://gotax.bioinf.mpi-inf.mpg.de`). Furthermore, there is a SOAP server available for integration of GOTax into other services. FSST is an implementation of the functional similarity measures, which allows for using customized sets of annotated gene products. To investigate mechanisms of pathogenicity and new drug target candidates, we have applied GOTax in a comparison between pathogenic *Mycobacteria* and mammals [3]. This comparison revealed different molecular functions present in *Mycobacteria* and absent in mammals. One example is the "UDP-N-acetylmuramate dehydrogenase activity", which is one step in the synthesis of bacterial peptidoglycan, and proteins with this function are candidates for drug targets.

We used the GOTax platform for an analysis of the functional space of Pfam families. We computed all pairwise molecular function similarities between Pfam families and performed multidimensional scaling to create a two-dimensional map of the Pfam functional space (Figure 14.13). In this map, Pfam families annotated with the same function form well-defined clusters. Furthermore, Pfam families with more general functional annotations locate towards the outer rim of the map.

A second application of our functional similarity measures is the area of domain-domain interactions (see also Section 14.6.2). Different bioinformatics methods for predicting domain-domain interactions (DDIs) have been developed. Although these computational methods

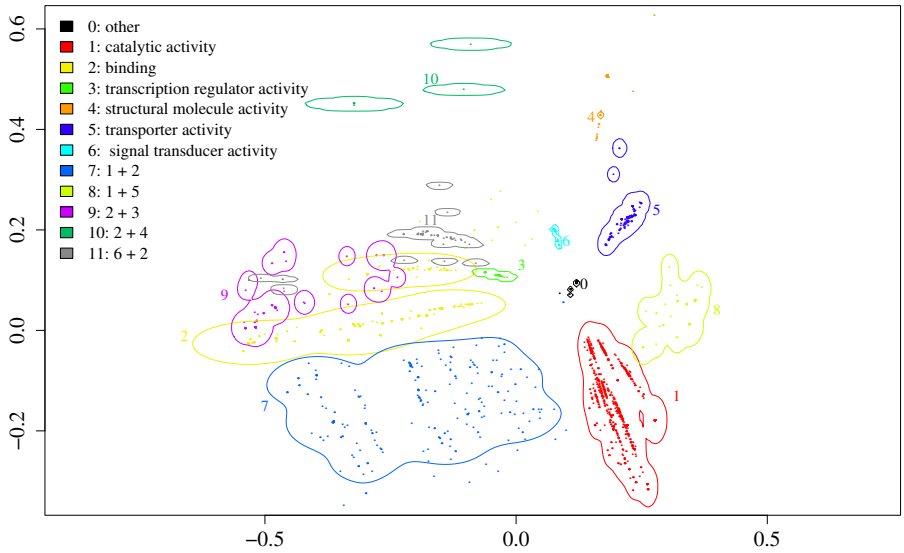


Figure 14.13: Map of the functional space of Pfam families. The colors were chosen according to the molecular function annotation of the Pfam families.

provide confidence scores, further evaluation of the predictions is required. Therefore, we assessed the reliability of three predicted DDI datasets using the $funSim$ measure [4]. In case of two of the predicted sets, our analysis shows that the functional similarity rises with increasing confidence score. This compares well with the corresponding increase of the validation rate of predicted interactions by structurally derived DDIs taken from iPfam. Based on our results, we could define $funSim$ cut-offs for high-confidence DDI predictions. This is important for improving the prediction of protein-protein interactions based on DDIs and for assessing the reliability of PPIs derived by high-throughput experiments.

References

[1] D. Lin. An information-theoretic definition of similarity. In *Proc 15th Int'l Conf. on Machine Learning (ICML-98)*, 1998.

[2] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proc 14th Int'l Joint Conf. Artifical Intelligence*, 1995, pp. 448–453.

- [3] 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, June 2006.

- [4] A. Schlicker, C. Huthmacher, F. Ramírez, T. Lengauer, and M. Albrecht. Functional evaluation of domain-domain interactions and human protein interaction networks. In D. Huson, O. Kohlbacher, A. Lupas, K. Nieselt, and A. Zell, eds., *German Conference on Bioinformatics (GCB 2006)*, Tübingen, Germany, September 2006, *Lecture Notes in Informatics*, vol. P-83, pp. 115–126. Gesellschaft für Informatik.

- [5] 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, March 2007.

14.6.2 Evaluation of Human Protein Interaction Data

Investigators: Fidel Ramírez and Mario Albrecht

Protein-protein interactions (PPIs) are fundamental for almost all cellular processes. The identification of PPIs reveals functional relationships between proteins and uncovers proteins involved in human diseases. Currently, the complete human interactome is estimated to contain 200,000 to 400,000 interactions [3]. In contrast, only about 40,000 experimentally verified interactions are known [2]. Most of these interactions are the result of small-scale experiments, but two high-throughput initiatives have recently provided almost 6,000 new interactions using the yeast two-hybrid technique [5, 9]. Since the experimental coverage of the human interactome is still low, various bioinformatics methods have been developed to predict PPIs.

We designed a relational database to integrate the biological information on human PPIs contained in diverse datasets. Currently our database contains 451,039 human interactions between 16,317 gene products. Only 9.12% of the interactions have been obtained by experiments, the rest by prediction methods. We also developed the NetworkAnalyzer plugin (`http://med.bioinf.mpi-inf.mpg.de/netanalyzer/`) for Cytoscape that computes topological network parameters such as degree distributions, diameters, shortest path lengths, and clustering coefficients. Another plug-in named DomainNetworkBuilder presented at the ECCB in 2005 decomposes interacting proteins into their respective domains and computes a putative network of corresponding domain-domain interactions [1].

To gain insight into the value of the publicly available human interaction data, we compared predicted datasets, high-throughput results, and manually curated PPIs taken from the literature [4]. Such an evaluation is not only important for further methodological improvements, but also for increasing the confidence in functional hypotheses derived from predictions. We assessed the quality and the potential bias of the different datasets using a new functional similarity measure based on the Gene Ontology [8] (see also Section 14.6.1), structural iPfam domain-domain interactions, likelihood ratios, and topological network parameters.

Our analysis revealed major differences between predicted datasets, but some of them also scored at least as highly as the experimental datasets regarding multiple quality measures (Figure 14.14). Therefore, since only small pairwise overlap between most datasets is observed, they may be combined to enlarge the available human interactome data. For this purpose, we additionally studied the influence of protein length on data quality and the number of disease proteins covered by each dataset. We could further demonstrate that protein interactions predicted by more than one method achieve an elevated reliability. We also find that experimentally derived interactions are prone to false positives, but, with the use of appropriate filters, subsets of elevated confidence can be derived.

Recently, we have used our whole compiled set of experimentally derived and predicted PPIs in collaborations within the European Network of Excellence BioSapiens and with a biological research group headed by Professor Russell Middaugh, Department of Pharmaceutical Chemistry, University of Kansas, USA. For the latter American researchers, we contributed detailed network analyses to their work on yeast and human proteins binding five cellular polyanions (actin, tubulin, heparin, heparan sulfate, and DNA). They used protein microarrays for approximately 4,000 yeast and 5,000 human proteins in an attempt to

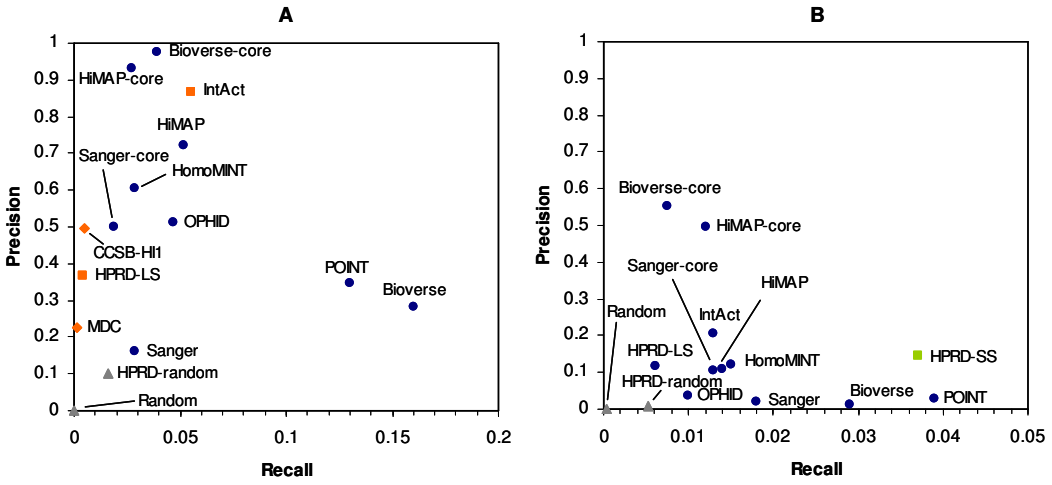


Figure 14.14: Recall-precision plots using (A) the manually curated human protein reference database (HPRD) or (B) two combined yeast two-hybrid datasets as positive reference sets. Circles denote predicted datasets of PPIs, diamonds high-throughput data, squares manually curated interactions taken from the literature, and triangles randomized interaction datasets.

improve the understanding of the functional characteristics of polyanion-protein interactions in yeast [7] and human cells [6].

For our BioSapiens partners, in the context of a comprehensive joint investigation of the impact of alternative splicing on ENCODE gene products, we studied the interactions formed by the ENCODE protein complement, regarding network topology, molecular function, biological process, domains, and medical relevance [10]. Additionally, domain-domain interactions responsible for specific interactions were suggested based on confidence values. A web server based on the Distributed Annotation System (DAS) was implemented to provide our interaction data as additional annotation to ENCODE genes. Currently, we are extending the DAS protocol to share and annotate different types of molecular interactions in a client-server model. This aims at promoting the exchange and utilization of new interaction datasets released by different research groups worldwide.

References

- [1] M. Albrecht, C. Huthmacher, S. C. Tosatto, and T. Lengauer. Decomposing protein networks into domain-domain interactions. *Bioinformatics*, 21(Suppl. 2):ii220–ii221, 2005.

[2] T. K. B. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, S. S. Mohan, S. Sharma, S. Pinkert, S. Nagaraju, B. Periaswamy, G. Mishra, K. Nandakumar, B. Shen, N. Deshpande, R. Nayak, M. Sarker, J. D. Boeke, G. Parmigiani, J. Schultz, J. S. Bader, and A. Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–293, Mar 2006.

[3] G. T. Hart, A. K. Ramani, and E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120, 2006.

[4] F. Ramírez, A. Schlicker, Y. Assenov, T. Lengauer, and M. Albrecht. Computational analysis of human protein interaction networks. *submitted*, 2007.

[5] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005.

• [6] N. Salamat-Miller, J. Fang, C. W. Seidel, Y. Assenov, M. Albrecht, and R. C. Middaugh. A network-based analysis of polyanion-binding proteins utilizing human protein arrays. *Journal of Biological Chemistry*, 282(14):10153–10163, 2007.

• [7] N. Salamat-Miller, J. Fang, C. W. Seidel, A. M. Smalter, Y. Assenov, M. Albrecht, and C. R. Middaugh. A network-based analysis of polyanion-binding proteins utilizing yeast protein arrays. *Molecular and Cellular Proteomics*, 5(12):2263–2278, 2006.

[8] A. Schlicker, C. Huthmacher, F. Ramírez, T. Lengauer, and M. Albrecht. Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23:in press, 2007.

[9] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122:957–68, 2005.

• [10] 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, P. László, 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 of the United States of America*, 104(13):5495–5500, 2007.

14.6.3 Detailed Human Protein Analyses

Investigators: Gabriele Mayr, Christoph Welsch, and Mario Albrecht

Introduction

We aim at the development of bioinformatics methods for advancing the understanding of disease processes and of the effect of genetic variations on protein function and drug therapies. To this end, application studies on structure and function prediction of medically relevant proteins are being conducted in cooperation with research groups from medical institutes in the context of our projects in the German National Genome Research Network (NGFN) and the Clinical Research Group on hepatitis C (funded by the German Research Foundation DFG) as well as in the European Networks of Excellence named BioSapiens (bioinformatics) and VIRGIL (hepatitis C). Our computational findings have already led to

plausible biological hypotheses for prioritizing further experiments. Meanwhile, several experiments have been completed successfully and have advanced the molecular understanding of impaired cellular mechanisms involved in different disorders. In the following, we detail our bioinformatics studies.

Crohn disease and resistance proteins

The autoinflammatory Crohn disease is in the focus of our medical cooperation partners from the Christian-Albrechts University in Kiel, Germany (Prof. Dr. Stefan Schreiber, Institute for Clinical Molecular Biology). Recently, a genome- wide association study of 735 patients with this bowel disease implicated a new SNP in the ATG16L1 gene with higher disease susceptibility [7]. ATG16L1 is a critical component of the cellular phagosome pathway, which is essential for the degradation of intracellular bacteria by immune cells. To study the structural impact of the identified SNP encoding the amino acid change T300A on the protein level, we assembled a multiple sequence alignment including all ATG16L1 family members and remotely related proteins with known 3D structures. Our sequence alignment demonstrated that the variant position is well conserved and located within a WD40-repeat domain. We also constructed a three-dimensional protein structure model of this domain (Figure 14.15). Its propeller-like fold due to WD40 repeats is found to be involved in protein-protein interactions in a variety of proteins. Thus, in case of ATG16L1, a mutated WD40-repeat domain may impair interactions crucial for proper phagosome functioning. This may subsequently cause inflammation due to an insufficient degradation of bacteria infiltrating human cells from the bowel. Therefore, our results support the emerging notion of Crohn disease as an intestinal inflammatory barrier disorder. This is also supported by the discovery of further candidate genes using cDNA microarrays [4].

Furthermore, guard proteins, so-called NLRs, are also associated with autoinflammatory diseases such as Crohn disease [13]. NLRs are important intracellular receptors of the innate immune system for detecting pathogens [13]. Since NLRs are evolutionary closely related to plant resistance proteins, an exchange of knowledge about molecular mechanisms between mammals and plants is feasible. To stimulate research in this area further, we have teamed up with plant experts (Dr. Frank Takken, Institute for Life Sciences, University of Amsterdam, The Netherlands, and Dr. Wladimir Tameling, The Sainsbury Laboratory, John Innes Centre, United Kingdom). Three joint articles provided one of the first structure-based views and mechanism interpretations of resistance protein domains and the dysfunctional effect of amino acid mutations on plant immunity [1, 15, 14]. The impact of this work is reflected in subsequent reviews on the plant immune system in prominent journals citing our structural studies [6, 8, 5].

Hepatitis C

Hepatitis C is a viral, mostly chronic, disease, which currently affects half a million patients in Germany and more than 170 million people world-wide. Infection with the hepatitis C virus (HCV) is the major cause for liver diseases such as cirrhosis and liver cancer. Therapies against Hepatitis C are so far effective only in up to 60% of all cases. Viral mutagenicity and subsequent drug resistance have become an important issue for treatment success. Our

department participates in an interdisciplinary clinical research group funded by the DFG in cooperation with the Saarland University Hospital, Homburg, Germany (Prof. Dr. Stefan Zeuzem, Department of Internal Medicine II). Experimentally, the study of virus infection is commonly realized with the HCV replicon system. In this context, we provide bioinformatics analysis of the protein sequence and structure of HCV proteins to study replication enhancing mutations. First results indicate an association of a subset of these mutations with a slower decrease of HCV RNA concentration during IFN-γ therapy in 5 out of 26 patients [12].

Further studies focused on the relevance of HCV p7 amino acid variations for the response to antiviral therapy with the drug amantadine. The structural protein p7 is a cation-selective ion channel in the membrane of the endoplasmatic reticulum (ER) essential for viral replication. Amantadine is an ion channel inhibitor used in combination with interferon and ribavirin for the treatment of HCV genotype 1 infected patients. We investigated 86 patients with chronic HCV infection for associations between antiviral therapy response and p7 amino acid mutations. The resulting HCV p7 sequence variations were evaluated with respect to channel function, structure and interaction with amantadine [9]. Our helical-wheel model of the HCV p7 ion channel structure (Figure 14.15) locates amino acids of potential importance for channel function. Based on this model, we could suggest that substitutions such as L20F may lead to a structural change in the channel pore, impeding the interaction of amantadine with the HCV p7 channel and thus decreasing the response to antiviral therapy.

Furthermore, we characterized structural and functional properties of the non-structural protein 4B of HCV and other members of the flaviviridae family [17]. We identified a basic leucine zipper (bZIP) motif in the N-terminal part of NS4B, which may be involved in a physical interaction with CREB-RP/ATF6beta and the modification of the cellular ER-stress response after viral infection. Our findings were used to investigate NS4B mutational sites concerning their impact on viral kinetics and interferon response [16].

Neurodegenerative disorders

We continued our previous studies of neurodegenerative diseases in cooperation with biological groups from the Max Planck Institute for Genetics, and the Max Delbruck Center for Molecular Medicine, Berlin, Germany (Dr. Sylvia Krobitsch and Prof. Dr. Erich Wanker). The evolutionary unrelated proteins ataxin-2 and ataxin-3 are the products of genes causative of spinocerebellar ataxia type 2 and 3 (SCA2 and SCA3), respectively. Both contain a polyglutamine tract encoded by CAG repeats, whose expansion beyond a certain threshold causes the associated autosomal-dominantly inherited disorders. The diseases SCA2 and SCA3 belong to a heterogeneous group of trinucleotide repeat disorders, which includes Huntington's disease and several other spinocerebellar ataxia types such as SCA1, SCA7 and SCA10. The disorders share common phenotypical features such as the progressive degeneration of specific vulnerable neuron populations. Clinical main features are ataxia and dementia, which can also resemble parkinsonism. However, the expression of the disease-associated genes occurs in a great variety of tissues and is not restricted to neuronal cells.

In collaborative work, we could demonstrate that the protein ataxin-2, causative of the spinocerebellar ataxia type 2 (SCA2), interacts with two members of the endophilin family, endophilin-A1 and endophilin-A3. In particular, our predicted interaction sites of the endophilins with ataxin-2 could be confirmed experimentally [11]. Since endophilin-A3 also

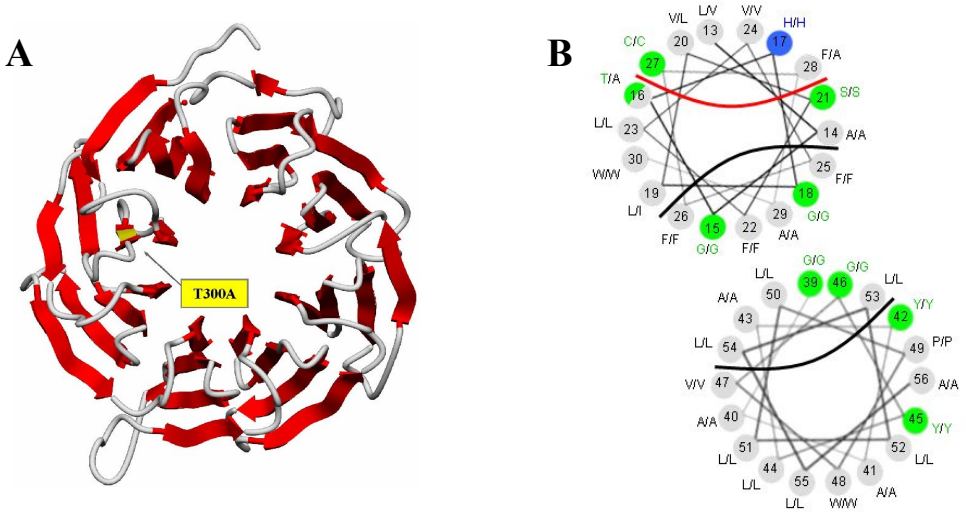


Figure 14.15: (A) 3D structure model of the WD40-repeat domain in the ATG16L1 protein. The location of the disease-associated SNP position is indicated in yellow. (B) Helical wheel model of one HCV p7 protomer showing two transmembrane helices TM1 and TM2. Apolar amino acids are colored gray, polar amino acids green, and the position of a basic histidine is blue.

interacts with huntingtin underlying Huntington's disease, both proteins ataxin-2 and huntingtin may work in related pathways. Moreover, we analyzed the binding site of the ATPase p97, a molecular chaperone, to the ubiquitin protease ataxin-3 [2]. The site consists of a highly conserved motif of arginines and lysines located close to the polyglutamine track of ataxin-3. This finding was surprising because such a motif typically functions as targeting signal for the translocation of proteins to the nucleus, which is not the case here. However, we also discovered that a similar stretch of four basic amino acids in the ubiquitin chain assembly factor E4B is critical for binding p97. Overall, our results defined the interaction of ataxin-3 with p97 as a potential target for therapeutic intervention. Apart from that, we have established a new contact with Prof. Dr. Christopher Ross (Division of Neurobiology, Johns Hopkins University School of Medicine) for analyzing protein networks not only for ataxia, but also for Parkinson's disease, because associated proteins appear to work in similar molecular processes.

Further protein analyses

The heterodimer MutLα of MLH1 and PMS2 plays a central role in human DNA mismatch repair. It interacts ATP-dependently with the mismatch detector MutSα and assembles and controls further repair enzymes. Our cooperation partner Dr. Guido Plotz at the Saarland University Hospital, Homburg, Germany, mutated amino acids to test if the interaction of MutLα with DNA-bound MutSα is impaired by cancer-associated mutations in MLH1. He identified four amino acid changes that abolished interaction as well as mismatch repair activity. Homology modeling of MLH1 showed that these amino acids clustered in a small

accessible surface patch, suggesting that the major interaction interface of MutLα for MutSα is located on the edge of a β-sheet that backs the MLH1 ATP-binding pocket [10]. We also found that this patch corresponds to a conserved potential protein–protein interaction interface present in both human MLH1 and its bacterial homologue MutL. Site-specific cross-linking of MutL to MutS from this patch verified our prediction because the bacterial MutL–MutS complex was established by the corresponding interface in MutL.

Another interesting application study conducted within an international cooperation group of European and American researchers was the comprehensive bioinformatics characterization of a new selenoprotein family named SelJ bearing strong homology to the large family of ADP-ribosylation enzymes [3]. Here, we studied the active site of the only available enzyme structure in detail, which had remained completely uncharacterized as an output of a structural genomics project. We were then able to transfer our functional observations to the closely related SelJ homologs and proposed further experiments to uncover the as yet unknown molecular function of the SelJ family.

References

- [1] M. Albrecht and F. L. W. Takken. Update on the domain architectures of NLRs and R proteins. *Biochemical and Biophysical Research Communications*, 339(2):459–462, 2006.
- [2] A. Boeddrich, S. Gaumer, A. Haacke, N. Tzvetkov, M. Albrecht, B. O. Evert, E. C. Müller, R. Lurz, P. Breuer, N. Schugardt, S. Plaßmann, K. Xu, J. M. Warrick, J. Suopanki, U. Wüllner, R. Frank, U. F. Hartl, N. M. Bonini, and E. E. Wanker. An arginine/lysine-rich motif is crucial for VCP/p97-mediated modulation of ataxin-3 fibrillogenesis. *EMBO Journal*, 25(7):1547–1558, 2006.
- [3] S. Castellano, A. V. Lobanov, C. Chapple, S. V. Novoselov, M. Albrecht, D. Hua, A. Lescure, T. Lengauer, A. Krol, V. N. Gladyshev, and R. Guigó. Diversity and functional plasticity of eukaryotic selenoproteins: Identification and characterization of the SelJ family. *Proceedings of the National Academy of Sciences*, 102(45):16188–16193, 2005.
- [4] C. M. Costello, N. Mah, R. Häsler, P. Rosenstiel, G. H. Waetzig, A. Hahn, T. Lu, Y. Gurbuz, S. Nikolaus, M. Albrecht, J. Hampe, R. Lucius, G. Klöppel, H. Eickhoff, H. Lehrach, T. Lengauer, and S. Schreiber. Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays identifies novel candidate disease genes. *PLoS Medicine*, 2(8):771–787, 2005.

  [5] J. L. Dangl. Plant science. nibbling at the plant cell nucleus. *Science*, 315(5815):1088–1089, Feb 2007.

  [6] B. J. DeYoung and R. W. Innes. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat Immunol*, 7(12):1243–1249, Dec 2006.
- [7] 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.

  [8] J. D. G. Jones and J. L. Dangl. The plant immune system. *Nature*, 444(7117):323–329, Nov 2006.
- [9] U. Mihm, N. Grigorian, C. Welsch, E. Herrmann, B. Kronenberger, G. Teuber, M. von Wagner, W.-P. Hofmann, M. Albrecht, T. Lengauer, S. Zeuzem, and C. Sarrazin. Amino acid variations

in hepatitis C virus p7 and sensitivity to antiviral combination therapy with amantadine in chronic hepatitis C. *Antiviral Therapy*, 11(4):507–519, 2006.

- [10] G. Plotz, C. Welsch, L. Giron-Monzon, P. Friedhoff, M. Albrecht, A. Piiper, R. M. Biondi, T. Lengauer, S. Zeuzem, and J. Raedle. Mutations in the MutSalpha interaction interface of MLH1 can abolish DNA mismatch repair. *Nucleic Acids Research*, 34(22):6574–6586, 2006.
- [11] M. Ralser, U. Nonhoff, M. Albrecht, T. Lengauer, E. E. Wanker, H. Lehrach, and S. Krobitsch. Ataxin-2 and huntingtin interact with endophilin-A complexes to function in plastin-associated pathways. *Human Molecular Genetics*, 14(19):2893–2909, 2005.
- [12] C. Sarrazin, U. Mihm, E. Herrmann, C. Welsch, M. Albrecht, U. Sarrazin, S. Traver, T. Lengauer, and S. Zeuzem. Clinical significance of in vitro replication-enhancing mutations of the hepatitis C virus (hcv) replicon in patients with chronic HCV infection. *The Journal of Infectious Diseases*, 192(10):1710–1719, 2005.
- [13] S. Schreiber, P. Rosenstiel, M. Albrecht, J. Hampe, and M. Krawczak. Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nature Reviews Genetics*, 6(5):376–388, 2005.
- [14] F. L. Takken, M. Albrecht, and W. I. Tameling. Resistance proteins: molecular switches of plant defence. *Current Opinion in Plant Biology*, 9(4):383–390, 2006.
- [15] W. I. Tameling, J. H. Vossen, M. Albrecht, T. Lengauer, J. A. Berden, M. A. Haring, B. J. Cornelissen, and F. L. W. Takken. Mutations in the NB-ARC domain of I-2 that impair ATP hydrolysis cause autoactivation. *Plant Physiology*, 140(4):1233–1245, 2006.
- [16] M. W. Welker, W.-P. Hofmann, C. Welsch, M. von Wagner, E. Herrmann, T. Lengauer, S. Zeuzem, and C. Sarrazin. Correlation of nonstructural (NS) 4B amino acid variations with initial viral kinetics during interferon alfa-based therapy in HCV 1b-infected patients. *Journal of Viral Hepatitis*, 14, 2007.
- [17] C. Welsch, M. Albrecht, J. Maydt, E. Herrmann, M. W. Welker, C. Sarrazin, A. Scheidig, T. Lengauer, and S. Zeuzem. Structural and functional comparison of the non-structural protein 4B in flaviviridae. *Journal of Molecular Graphics and Modelling*, 25, 2007.

14.7 Computational Genomics with Applications to Cancer

In today's mass medicine, the diagnosis of diseases and the prognosis of disease progression are typically based on clinical and histopathological measurements. However, often patients with similar clinical characteristics can react differently to the same therapy. On average, only two thirds of all patients react as desired on the given therapy, for the rest the dose is too low or the patients show undesired side effects.

There are two main factors for the different disease courses and drug reactions. One reason is the different underlying genetic predisposition, the second is the inherent variability even in the presence of similar clinical as well as genetic measurements. We develop and apply statistical methodology for dealing with these problems. Our main goal is to better characterize patient subgroups based on genomic measurements which can enable a so-called *personalized medicine*.

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 around 30.000 genes can be measured simultaneously. We develop methods that integrate biological knowledge about gene functions in order to reduce the complexity of explanatory

genetic models, see Subsection 14.7.1. 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. We also developed models for genetic tumor progression that were successfully applied for characterizing patients subgroups with different expected clinical outcome, see Subsection 14.7.2. These models make use of CGH (comparative genomic hybridization) data that describe gains and losses of parts of chromosomes in the tumor cells. We extend this approach to arrayCGH data that also contain copy number changes in tumor cells, but with extremely higher resolution, see Subsection 14.7.3 for details.

Furthermore, we recently put much effort into better understanding of non-genetic causes of cancer, which comprises aberrant DNA methylation, chromatin structure and other epigenetic modifications. We developed methods and prototyped a software infrastructure for the emerging field of computational epigenetics, addressing the following topics (see Subsection 14.7.4 for details): Prediction of epigenetic modifications, refinement of the pivotal CpG island definition, automated curation of experimental DNA methylation data and optimization of epigenetic biomarkers for clinical use.

14.7.1 Transcriptomics

Investigators: Adrian Alexa and Jörg Rahnenführer

The result of a typical microarray experiment is a long list of genes with corresponding expression measurements. This list is the starting point for a meaningful biological interpretation. A major challenge for transcriptomics is to identify important biological processes or functions from gene expression data by scoring the relevance of predefined functional gene groups, for example based on Gene Ontology (GO) [3]. This type of analysis is known as gene set enrichment analysis (GSEA) or in a more broader scope as group testing [5].

Scoring the Relevance of Gene Ontology Terms

We have developed methods that increase the explanatory power of gene set enrichment by integrating knowledge about similarities and dependencies between the gene sets into the calculation of the statistical significance [1]. The intuitive and simple to interpret method, called *elim*, iteratively removes the genes mapped to significant GO terms from more general (higher level) GO terms. A more complex algorithm, called *weight*, was introduced to smooth out the elimination of genes. Here genes annotated to a GO term receive weights based on the scores of neighboring GO terms. Both methods analyze the GO graph structure, localize dependencies between GO terms, and remove them. This means that two neighboring GO terms that receive a significant score exhibit enrichment with different genes. Both algorithms can accommodate a large class of group based test statistics, making them suitable for a broader range of biological problems.

We have evaluated both algorithms on various cancer microarray datasets [1, 4]. Due to the subjectiveness in interpreting the results obtained on real datasets we introduce a novel evaluation scheme in which selected GO terms are artificially enriched, and the performance of the methods is quantified with respect to the number of correctly identified enriched terms. In comparison with state-of-the-art methods for scoring functional terms, the new

algorithms point at additional areas in the GO graph with significant biological processes or functions.

To provide access to the novel methods to the bioinformatic community we developed the `topGO` software. The package has been written in the statistical language R and is part of the Bioconductor repository [2]. `topGO` allows for semi-automated enrichment analysis. The data analysis pipeline consists of normalizing the arrays and inference of gene expression values, gene-wise correlation analysis with a phenotype of interest, gene set enrichment analysis, and interpretation and visualization of the results.

Applications to cancer data

Prostate cancer is the second most common type of cancer in men with up to 40% of elderly men developing prostate carcinomas. Recent research in molecular cancer has focused on classifying prostate cancers and on identifying targets for novel therapies.

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 [4]. A multivariate linear model was used to estimate the main effects of alterations of chromosome 8 and DNA hypomethylation and the interaction effect of these two factors. This gene-wise analysis revealed that factors like alterations of chromosome 8 and DNA hypomethylation in prostate cancer do not cause each other, but rather converge during progression. Enrichment analysis performed with the `topGO` software identifies interesting biological processes related to innate immunity, cytoskeletal organization and cell adhesion as common targets of both alterations. Close investigation of significantly enriched GO terms revealed additional genes which are strongly related to prostate cancer progression but were not identified as highly significant for the gene-wise interaction effect. New candidate genes for cancer prognosis were therefore selected from the top significant GO terms for further investigations.

The study thus highlights novel mechanisms in prostate cancer progression and identifies novel candidate genes for diagnostic and therapeutic purposes.

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. 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.

- [4] 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*, 6:14, 2007.

[5] Subramanian, A., *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, October 25 2005.

14.7.2 Genetic Tumor Progression Models

Investigators: Jörg Rahnenführer and Jasmina Bogojeska

In cancer research, prediction of time to death or relapse is important for a meaningful tumor classification and selecting appropriate therapies. There is increasing interest in identifying genetic markers that better capture the status of a tumor than classical clinical markers. The accumulation of genetic alterations during tumor progression can be used for the assessment of genetic tumor progression.

Genetic Progression Models

Previously, we introduced tree mixture models that can be used for modelling genetic progression as multiple diverse pathogenetic routes [1]. In a single tree model, genetic events are assumed to be non-reversible, thus the disease process can be fully described by the accumulation of genetic aberrations. In the mixture model, more than one tree component is estimated. Every model component describes the disease process for a subset of observations (tumors). We proposed an EM-like algorithm for estimating such tree mixture models from cross-sectional data [1]. A corresponding C software package contains functions for model estimation, validation, and simulation [2]. Based on these programs, we implemented an easy to use, flexible, open access package written in the statistical programming language R that includes additional functionality [3].

The EM-like algorithm for fitting mixture models assumes that the number of trees K is known. For large data sets, estimating this model parameter in a cross-validation framework, as proposed in [1], becomes computationally infeasible. We developed a new model selection criterion that is a modification of the BIC (Bayesian Information Criterion). It incorporates a similarity measure for estimating the structural redundancy between tree components in the penalization term of the standard BIC. Experimental results on simulated and on real data show that most classical model selection criteria tend to select models with far too many tree components. Only cross-validation and the modified BIC recover the correct number of trees most of the time. However, at the same optimal performance, the runtime of the new BIC modification is about one order of magnitude lower.

From the tree mixture models a genetic progression score (GPS) can be derived that estimates the genetic status of a tumor. Using Cox regression models we previously demonstrated that the GPS is a medically relevant prognostic factor. For several cancer types, the GPS can be used to discriminate between patient subgroups with different clinical outcome [5]. We have extended our models for estimating tumor progression to changes on gene level, based on arrayCGH measurements. This enables a more precise determination of relevant genomic regions, see Section 14.7.3 for details.

GPS values were shown to have predictive power for estimating drug resistance in HIV or the survival time in cancer. Still, the reliability of the exact values of such complex markers derived from graphical models can be questioned. In a simulation study, we analyzed various aspects of the stability of estimated mutagenetic trees mixture models. We show that the

induced probabilistic distributions and the tree topologies are recovered with high precision by an EM-like learning algorithm. However, only for models with only one major model component, even GPS values for single patients can be reliably estimated [3].

Progression in Meningiomas

Meningiomas are mostly benign tumors that originate from the coverings of brain and spinal cord. We calculated tree mixture models for a dataset containing cytogenetic measurements from 661 meningioma patients. The estimated progression model agrees with *a priori* medical knowledge and was used to assign GPS values to given individual tumors. Fitting multivariate Cox regression models we confirmed a high correlation between large GPS values and early tumor recurrence (significance level $p < 10^{-4}$). The information gain due to GPS holds true after adjusting for gender and age at diagnosis. Tumor location was also shown to be associated with different levels of cytogenetic progression.

In a second study we analyzed a small group of meningiomas with the atypical property that in the tumor cells chromosomal regions are more often gained than lost. The tumors were shown to constitute a clinically relevant entity of biologically aggressive meningiomas [4].

Response to Chemotherapy for Glioblastomas

We investigated the relevance of genetic aberrations in tumor cells of glioblastoma patients for response to temozolomide chemotherapy (TMZ). Multivariate Cox regression models including interaction effects were used to explore the relationships between patient survival and explanatory variables such as genomic alterations and type of therapy. For patients without adjuvant therapy, deletions on chromosomes 9p and 10q indicated poorer survival. However, patients with these molecular alterations did particularly benefit from TMZ treatment [6].

References

- [1] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer. Learning multiple evolutionary pathways from cross-sectional data. *Journal of Computational Biology*, 12(6):584–598, 2005.
- [2] N. Beerenwinkel, J. Rahnenführer, R. Kaiser, D. Hoffmann, J. Selbig, and T. Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107, 2005.
- [3] J. Bogojeska. Stability analysis of oncogenetic trees mixture models. Masters thesis, Universität des Saarlandes, January 2007.
- [4] 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*, p. Published online January 17, 2007.
- [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] S. Wemmert, R. Ketter, J. Rahnenführer, N. Beerenwinkel, M. Strowitzki, W. Feiden, C. Hartmann, T. Lengauer, F. Stockhammer, K. D. Zang, E. Meese, W.-I. Steudel, A. von Deimling, and

S. Urbschat. Patients with high grade gliomas harboring delections of chromosomes 9p and 10q benefit from temozolomide treatment. *Neoplasia*, 7(10):883–893, 2005.

14.7.3 Analysis of CGH Data

Investigators: Laura Tolosi and Jörg Rahnenführer

Currently used clinical markers often prove to be insufficient for an accurate diagnosis and prognosis in cancer patients. Latest technological developments provide experimental information on genetic mutations observable in tumors, therefore allowing a more precise diagnosis. Array CGH technology measures with high resolution gene copy number imbalances in tumor cells, but the data often contains biological and experimental noise. Many statistical methods for the analysis of array CGH data are currently being proposed in the literature, with the general goal of extracting relevant information for tumor progression, tumor subtype or survival time prediction.

Our research is driven in particular towards finding evolutionary models that describe patterns of aberrations accumulation during tumor progression. The method we propose is summarized in three main steps: *Single-array aberrations detection*, *consensus region selection*, and *evolutionary model estimation*. We shortly describe our solutions for the first two steps, the last being subject to future work.

Single-array aberrations detection

The detection of aberrations in single CGH arrays is the most common problem in array CGH analysis, aimed at identifying chromosomal regions which are either gained or deleted. It typically involves smoothing as a preprocessing step in order to identify segments of constant copy number and then choosing a significance threshold in order to determine gains and losses. From the smoothing algorithms with proven good performance we choose GLAD [2], which resumes to fitting a piecewise constant function such that a penalized likelihood is maximized. In [4] we propose a new algorithm for determining aberration thresholds, based on the assumption that the distribution of the smoothed log-ratios is normal around zero with deviating tails. A robust normal is fitted to the smoothed log-ratios and significance thresholds at two standard deviations from the mean are computed.

Consensus region selection

A concurrent analysis of multiple CGH arrays from the same type of cancer can reveal common mutations which are considered key genetic events occurring during tumor development. The immediate consequence of identifying the set of relevant genomic mutations from an alignment of CGH arrays is pinpointing to genomic regions where oncogenes or tumor suppressor genes can be found. Despite the high biological interest in discovering tumor related genes, few algorithmic solutions have been proposed for consensus region selection [1]. Our innovation resides in a *segmentation algorithm*, which partitions the genome in regions of constant profile, therefore providing an exhaustive list of candidate regions, unlike most existing methods which consider only those regions with high aberration frequency. Keeping in mind that the events selected should improve current tumor stage or survival time

prediction, we propose to apply a classical *supervised feature selection* procedure to the set of regions given by the segmentation algorithm. The predicted variable can be any clinical marker.

Application to breast cancer

The array CGH analysis pipeline has been validated on a breast cancer data set consisting of 54 arrays described in [3]. The single array aberrations detection step identified gains and losses consistent with previous reports. The consensus region selection revealed an interesting gain event located in chromosome 8q which seems to be fairly accurate in predicting tumor grade and p53 status, two commonly used markers in breast cancer staging.

References

[1] Bergamaschi A, Kim YH, Wang P, Sørlie T, Hernandez-Boussard T, Lonning PE, Tibshirani R, Børresen-Dale AL, Pollack JR. Distinct Patterns of DNA Copy Number Alteration Are Associated with Different Clinicopathological Features and Gene-Expression Subtypes of Breast Cancer. *Breast cancer research*, 45:1033 – 1040, 2006.

[2] Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–22, 2004.

[3] Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Børresen-Dale AL, Brown PO. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *PNAS*, 99(20):12963–12968, 2002.

- [4] L. Tolosi. Analysis of array CGH data for the estimation of genetic tumor progression. Masters thesis, Universität des Saarlandes, June 2006.

14.7.4 Computational Epigenetics

Investigators: Christoph Bock and Konstantin Halachev

While the role of genetic aberrations (mutations and copy number changes) for cancer progression has been known for a long time, recent research highlights the importance of more volatile, epigenetic damages as a second root of cancer. Epigenetic regulation is an important factor controlling gene expression in eukaryote genomes. It comprises all heritable changes in gene function that cannot be explained by changes in the DNA sequence. Mechanistically, genes can be activated for transcription by tagging their promoters with specific histone modifications and they can be silenced by mechanisms such as DNA methylation or local condensation of chromatin structure. Errors in these regulatory processes knock out critical tumor suppressor genes as frequently as genetic aberrations do and are therefore an active area of research for both cancer diagnosis and treatment. Demand for bioinformatic methods and computational tools in the field of cancer epigenetics is rapidly increasing, due to complex experimental methods, increasingly large-scale analysis and the pressure to quickly translate scientific results into clinical practice. Our goal is to develop a bioinformatic methodology that addresses these issues and to implement a set of web services which provide powerful methods in a way that can be easily operated by the average bench biologist. We coined the term “computational epigenetics” [2] to summarize this research agenda.

Prediction of Epigenetic Modifications

Due to high costs for genome-wide experimental analysis, epigenome prediction is an attractive topic for bioinformatic research. For DNA methylation, which is the best understood epigenetic modification in mammals, we could – for the first time – show that prediction is possible with high accuracy, based on specific characteristics of the DNA sequence and its preferred structure [2]. We recently extended this finding to several other epigenetic modifications, and we applied our results in order to improve the mapping of CpG islands in the human genome (Bock et al. submitted). Our current work in this area aims at identifying characteristics of the DNA sequence that correlate with regions exhibiting aberrant epigenetic modifications in cancer.

Software Infrastructure for Computational Epigenetics Research

Due to the infancy of the field, few software tools exist to facilitate epigenetic research. Therefore, an important goal of our research is to implement bioinformatic methods for epigenetic research in easy-to-use software packages and make them available to the global research community. First, we developed a software tool that supports curation and low-level analysis of DNA methylation data derived by bisulfite sequencing. Our BiQ Analyzer provides an easy step-by-step analysis workflow to facilitate reproducible data analysis [1]. BiQ Analyzer is currently in use in several hundred labs worldwide and was selected by ABI to be part of the Applied Biosystems Software Community Program. Second, we are currently implementing a high-throughput, web-based service for epigenome analysis and prediction. A beta version of our EpiGRAPH software was released in 2006 and the public version will become available in summer 2007.

Cancer Biomarker Discovery and Optimization

Epigenetic modifications can provide important evidence for cancer detection and diagnosis. First, cancer cells often exhibit epigenetic damage long before a visible tumor develops. Second, histologically indistinguishable tumor types with different response to treatment can frequently be distinguished by their epigenetic patterns. For these reasons, the clinical use of epigenetic cancer biomarkers is anticipated to contribute significantly to improved cancer diagnosis and treatment. In cooperation with clinical researchers we have recently optimized and validated a DNA methylation biomarker that predicts chemotherapy resistance in glioblastomas (Mikeska, Bock, et al. The Journal of Molecular Diagnostics, in press). This biomarker is now in routine use at the National Brain Tumor Reference Center (University of Bonn), leading to more informed treatment decisions.

References

- [1] 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.
- [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.

14.8 Computational Chemical Biology

The work in this section focuses on docking and drug design and cheminformatics. 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 on 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 screening. 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. If there is no structural information available of the target protein, cheminformatics approaches like ligand screening can be used, which only consider the chemical properties of the ligand molecules. Our work focuses on the development of new and improved methods for settings for which the above algorithms do not perform well, aiming at both, biomolecular and chemical applications.

One major project in the field of biomolecular 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 (section 14.8.2 and 14.8.4). In a second project we are working on a combined bioinformatics/biophysics based algorithm for docking peptide and peptidic ligands into flexible binding sites (section 14.8.4). 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 immunoinformatics project in which a combined sequence- and structure-based method for the prediction of peptide binders to MHC-molecules was developed (section 14.8.3). 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 developed a new program, POEM (section 14.8.1), which combines the DOE algorithm with ensembles of different statistical learning methods for parameter optimization.

Next to the above projects which aim on biomolecular applications, there are two methodological projects in the field of cheminformatics. First, a new algorithm was developed for efficient docking of ligands into artificial receptors (section 14.8.7). Second, a new graph-based method is currently being developed for a ligand based screening of chemical space (section 14.8.6).

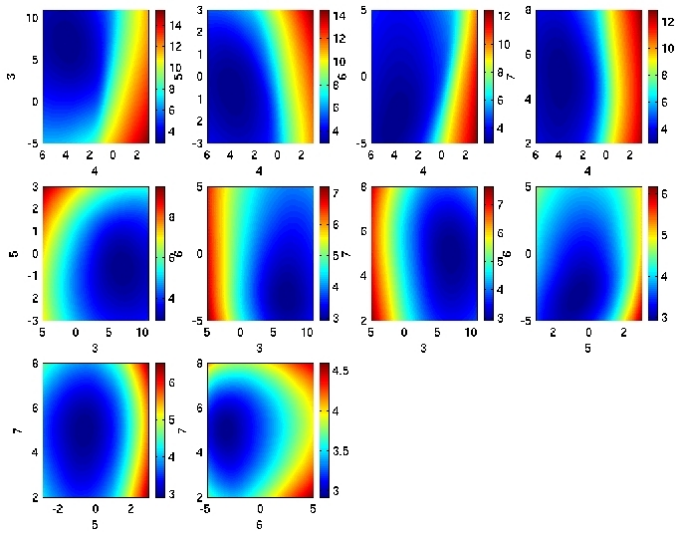


Figure 14.16: 2D Cuts through the calculated cost function landscape during the optimization of the FlexX scoring function for the kinase protein class.

14.8.1 Design and optimization of scoring functions

Investigator: Iris Antes

In bioinformatics processes like docking, binding, and folding are often described by simplified, empirical models. The development of a new computational model includes two major steps: namely, the design of its functional form and the adjustment of its parameters. Choosing the correct parameters is crucial for the quality and performance of the model. However, locating the best choices for the parameters can be a difficult task, depending on the complexity and roughness of the underlying parameter landscape. We developed a new method and program, POEM (Parameter Optimization using Ensemble Methods), for this purpose [1]. In POEM the DOE (Design Of Experiment) procedure [3] is combined with ensembles of different regression methods. The method consists of an iterative procedure that uses alternate evaluation and prediction steps leading to a very efficient search strategy in parameter space. For this purpose an approximate function is fitted to the cost function landscape at each cycle of optimization and the quality of this fit is improved from cycle to cycle by constantly augmenting the used data set. Ensembles of neural networks [2] and k-nearest neighbor methods were used for the fitting and a genetic algorithm was applied to locate the global minimum of the cost function landscape. The method was evaluated for the optimization of target specific scoring functions in molecular docking. The FlexX [4] and Screenscore [5] scoring function were optimized for the kinase and ATPase protein classes. The results are very promising: Starting from random parameters we are able to locate parameter sets which show superior performance to the original and other docking scoring functions. The optimization approach proved to be very efficient and converged very fast. The approximated cost function landscapes were very smooth, thus making the approach a promising method for optimizations on complex landscapes.

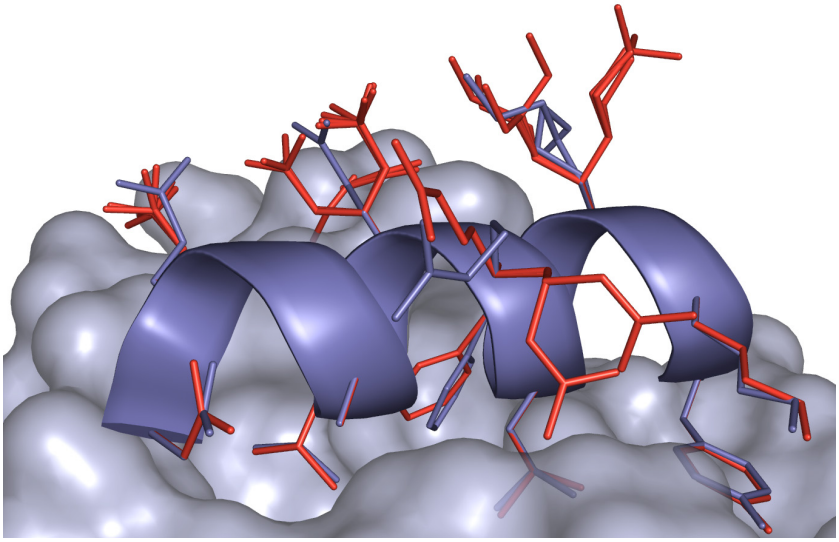


Figure 14.17: A helix (chain B, position 69-80) of human UDP-galactose 4-epimerase (PDB: 1EK6). The side chains predicted with IRECS are red, the side chains of the crystal structure are blue.

References

- [1] I. Antes, C. Merkwirth, and T. Lengauer. POEM: Parameter optimization using ensemble methods: Application to target specific scoring functions. *Journal of Chemical Information and Modeling*, 45(5):1291–1302, 2005.

[2] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[3] C. D. Montgomery. *Design and Analysis of Experiments*. Wiley, New York, 1991.

[4] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–489, Aug 1996.

[5] M. Stahl and M. Rarey. Detailed analysis of scoring functions for virtual screening. *J Med Chem*, 44(7):1035–42, Mar 2001.

14.8.2 Docking into Homology Models of Protein Structures

Investigators: Christoph Hartmann and Iris Antes

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 two major reasons for this are the limited accuracy especially of the predicted side-chain conformations in homology models and the dependency of the overall structure (backbone) of the model on the template structures

used. The latter can lead to severe inaccuracies especially at loop and hinge regions and binding sites. An additional refinement and validation of these structures is necessary before docking. For this purpose the programs DynaDock-Homo [6] and IRECS [3] were developed, the former improving the backbone and the latter the side-chain conformations of the model structures.

IRECS (Iterative REduction of Conformational Space) is a new tool for side-chain placement, which is especially tailored for 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. However, depending on the size of the binding site, this often leads to a combinatorial explosion of possible combinations of conformations. Thus it is crucial to preselect a small number of flexible side chains in the binding site, for which alternative conformations are used. IRECS provides such a set, which can directly be used in the FlexE [2] module of our docking software FlexX [5]. Using such an ensemble for each side chain, FlexE can identify the optimal side chain rotamers for each ligand pose efficiently with a mean field optimization algorithm, similar to the SCMF algorithm [4]. IRECS is driven by a new knowledge-based scoring function ROTA that we derived for scoring favorable side chain conformations. We evaluated our tool on a set of 160 crystal structures of proteins. Predicting only a single conformation per side chain, IRECS achieved a χ_1 accuracy of 84.7% and a χ_{1+2} accuracy of 74.3%, using a 40° cutoff. This is comparable to the widely used side-chain prediction tools SCWRL [1] and SCAP [7].

DynaDock-Homo allows for inhibitor-based refinement of homology models, using the additional information provided about the structure of the binding site by the structural scaffolds of known binders. For this purpose an iterative procedure was developed which consists of alternating docking and refinement steps. For the refinement a special simulation protocol was optimized which combines simulated annealing, molecular dynamics and Monte Carlo approaches. This strategy was successfully tested for several cytochrome P450 systems. An average reduction of the binding site backbone RMSD from 4.0 Å to 2.5 Å could be achieved, thus leading to models which are accurate enough for docking. In addition to the evaluation studies, the strategy was used during a drug design project searching for selective inhibitors of aldosterone synthase (CYP11B2) (see Section 14.8.5). The final models were successfully evaluated by checking their selectivity for known binders/non-binders through docking calculations.

References

[1] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):2001–2014, Sep 2003.

[2] H. Claussen, C. Buning, M. Rarey, and T. Lengauer. FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol*, 308(2):377–395, Apr 2001.

- [3] C. Hartmann, I. Antes, and T. Lengauer. IRECS: A new algorithm of the selection of most probable ensembles of side-chain conformations in protein models. *Protein Science*, 15, 2007.

[4] P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol*, 239(2):249–275, Jun 1994.

[5] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–489, Aug 1996.

- [6] M. Voets, I. Antes, C. Scherer, K. Biemel, C. Barassin, 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] Z. Xiang and B. Honig. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol*, 311(2):421–430, Aug 2001.

14.8.3 Structural Immunoinformatics

Investigator: Iris Antes

The binding of endogenous antigenic peptides to MHC class I molecules is an important step during the immunologic response of a host against a pathogen. Thus, various sequence- and structure-based prediction methods have been proposed for this purpose. The sequence-based methods are computationally efficient, but are hampered by the need of sufficient experimental data and do not provide a structural interpretation of their results. The structural methods are data-independent, but are quite time-consuming and thus not suited for screening of whole genomes. We developed new method, DynaPred, which performs sequence-based prediction by incorporating information obtained from molecular modeling. This allows us to perform large databases screening and to provide structural information of the results [1, 2].

DynaPred is a SVM-trained, quantitative matrix-based method for the prediction of MHC class I binding peptides, in which the features of the scoring matrix are energy terms retrieved from molecular dynamics simulations. At the same time we used the equilibrated structures obtained from the same simulations in a simple and efficient docking procedure. Our method consists of two steps: First, we predict potential binders from sequence data alone and second, we construct protein-peptide complexes for the predicted binders. So far, we tested our approach on the HLA-A0201 and HLA-B2705 allele. We constructed two prediction models, using local, position-dependent (DynaPredPOS) and global, position-independent (DynaPred) features. The former model outperformed the two sequence-based methods used in our evaluation; the latter shows a much higher generalizability towards other alleles than the position-dependent model. The constructed peptide structures can be refined within seconds to structures with an average backbone RMSD of 1.53 Å from the corresponding experimental structures.

References

- [1] I. Antes, 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.

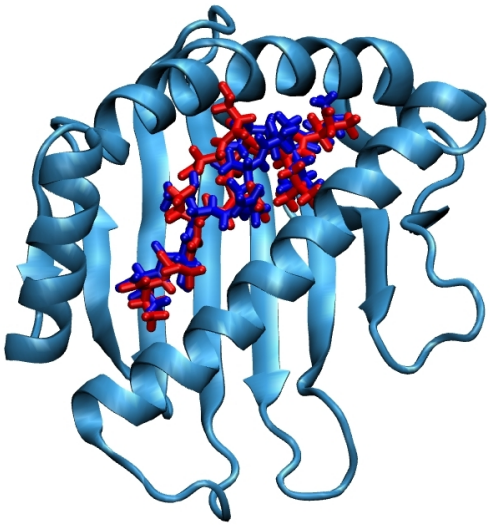


Figure 14.18: Experimental (blue) and docked (red) MHC-peptide complex (PDB ID: 1hhg)

- [2] W.-I. Siu. Computational prediction of MHC-peptide interactions. Masters thesis, Universität des Saarlandes, September 2005.

14.8.4 Protein-peptide Docking

Investigator: Iris Antes

To describe protein-ligand interactions realistically, it is necessary to account for the structural changes in both the receptor and the ligand during complex formation. In classical docking approaches based on discrete optimization algorithms this poses a very difficult problem due to the large number of possible side chain conformations at the receptors binding interface, often leading to a combinatorial explosion of possible combinations of conformations. The tool IRECS was developed to improve this situation for the docking of small ligands into closed binding sites (see Section 14.8.2). There are two situations, however, for which classical docking approaches do not work sufficiently well, namely the docking of large, highly flexible ligands, like peptides, and docking into flexible binding sites, for which induced fit phenomena are very important [2].

One of the most important settings for which the above restrictions are prohibitive is protein-peptide docking thus remaining one of the most difficult challenges in molecular docking. Many important biological phenomena involve specific molecular recognition mediated by protein-peptide interactions. Peptides also serve as natural inhibitors for proteins and as lead structures for drugs. Prominent examples of peptide based drugs are the inhibitors of the HIV and HCV–proteases and antibiotics inhibiting β-lactamases. There are two major obstacles for classical docking approaches when applied to peptide docking: First, the large number of rotatable bonds within peptides with more than 3-4 residues, and second, the fact that peptide binding is a collective process of all residues along the peptide chain, often leading to bound structures in which not all residues form optimal or strong bonds

with the protein. Most docking programs experience difficulties with the high flexibility and they assume a nearly optimal interaction pattern for each fragment of the ligand, which is often not the case.

Thus biophysical approaches like molecular dynamics are better suited for this purpose than discrete algorithms. Based on the experience from Sections 14.8.2, 14.8.3 and an application study, in which we applied our docking Program FlexX-Pharm to the docking of peptides with a length of 8-10 residues to the GYF-domain and compared the results with experimental structures of the peptide-protein complexes derived from NMR studies [1], a molecular dynamics based algorithm is currently being developed for the refinement of the poses of approximately placed highly flexible ligands in binding sites for which induced fit phenomena are important. First results obtained for the refinement of the peptide-MHC complexes from DynaPred are very promising.

References

- [1] W. Gu, M. Kofler, I. Antes, C. Freund, and V. Helms. Alternative binding modes of polyproline peptides binding to the GYF domain. *Biochemistry*, 44(17):6404–6415, April 2005.

[2] I. Trajkovski. Analysis of protein binding pocket flexibility. Masters thesis, Universität des Saarlandes, September 2004.

14.8.5 Docking Applications

Investigator: Iris Antes

Next to the methodological developments, substantial work was performed during several application projects in close collaboration with experimental groups:

The major project a classical drug design project and is 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 is to find inhibitors for aldosterone synthase, which 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 [3]. These compounds however, were not selective with respect to other steriod producing and metabolising 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 significantly other CYP P450s. Due to the difficulty of resolving membrane-binding proteins, there are no experimental structures available for the cytochromes CYP11B1 and CYP1 1B2. However, there is an intense ongoing effort to develop homology models of mammalian cytochrome P450s. Using the recently resolved human cytochrome CYP2C9 structure (PDB code: 1R9O) as template, we built 3D structural models for CYP11B2 and CYP11B1. Comparison of these models with two previously built structural models for CYP11B1 and CYP11B2 [1] 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

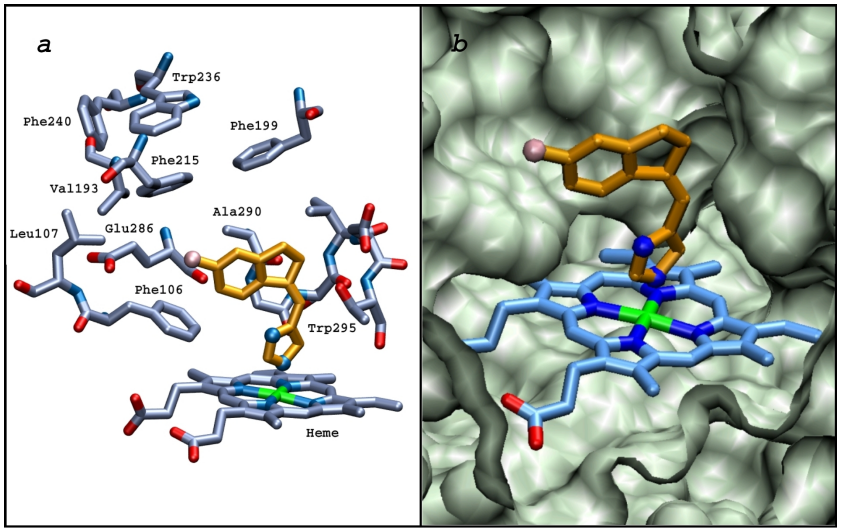


Figure 14.19: Structure of the CYP11B2-inhibitor complex: (a) details of the active side, (b) surface of the binding site surrounding the bound inhibitor and the heme-cofactor. The inhibitor is shown in brown.

templates used [5]. 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 DynaDock-Homo 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 help the lead refinement process by mapping the differences in the measured binding affinities for various newly synthesized compounds to the structural features of our docked complexes. In addition, the structural insights gained through the docking studies served as basis for the design of new compounds with higher binding affinities and selectivity [6, 7, 9, 8].

Next to the above project several application studies were performed to analyze the influence of mutations in the receptor on its binding capabilities (function). Mutations can have a major effect on the substrate and inhibitor binding affinity and the stability of the corresponding protein. Thus the prediction of these effects is of great importance in various fields from drug design to bioengineering. Based on the experiences and strategies described in the previous sections, the effect of binding site mutations on ligand binding were studied in several projects. Structural investigations were performed in two collaboration projects with the medical department of the University of the Saarland (Prof. S. Zeuzem and Dr. C. Welsch). In the first project we investigated the effect of a novel mutation in CYP21B which leads to 21-hydroxylase deficiency [2] . In the second project, molecular dynamics simulations and docking calculations were performed to explain the structural changes in the HCV-protease

binding site caused by resistance mutations (Welsch et al., in preparation). In collaboration with the biochemistry department of the Saarland University (Prof. R. Bernhardt) the effects of several mutations on the substrate binding in CYP106A2 were evaluated through docking and homology modeling with the goal of altering the regiospecificity of the protein by active-site engineering (Lisurek et al., submitted). In collaboration with the microbiology department of the Saarland University (Prof. F. Giffhorn and Dr. D. M. Heckmann-Pohl) combined docking and simulation studies are performed to decrease the specificity and increase the efficiency of pyranose-2 oxidase (P2Ox) [4]. P2Ox is a very important enzyme for biotechnological applications due to its ability to produce rare carbohydrates.

References

[1] N. V. Belinka, M. Lisurek, A. S. Ivanov, and R. Bernhardt. Modelling of three-dimensional structures of cytrochromes P450 11B1 and 11B2. *Journal of Inorganic Chemistry*, 87:197–207, 2001.

• [2] J. Bojunga, C. Welsch, I. Antes, M. Albrecht, T. Lengauer, and S. Zeuzem. Structural and functional analysis of a novel mutation of CYP21B in a heterozygote carrier of 21-hydroxylase deficiency. *Human Genetics*, 117(6):558–564, 2005.

[3] R. W. Hartmann, U. Müller, and P. B. Ehmer. Discovery of selective CYP11B2 (aldosterone synthase) inhibitors for the therapy of congestive heart failure and myocardial fibrosis. *Eur J Med Chem*, 38(4):363–6, Apr 2003.

• [4] D. M. Heckmann-Pohl, S. Bastian, S. Altmeier, and I. Antes. Improvement of the fungal enzyme pyranose 2-oxidase using protein engineering. *Journal of Biotechnology*, 124(1):26–40, 2006.

[5] S. B. Kirton, C. A. Kemp, N. P. Tomkinson, S. St-Gallay, and M. J. Sutcliffe. Impact of incorporating the 2C5 crystal structure into comparative models of cytochrome P450 2D6. *Proteins*, 49(2):216–31, Nov 2002.

• [6] 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, March 2005.

• [7] 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, April 2005.

• [8] M. Voets, I. Antes, C. Scherer, K. Biemel, C. Barassin, 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.

• [9] 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.

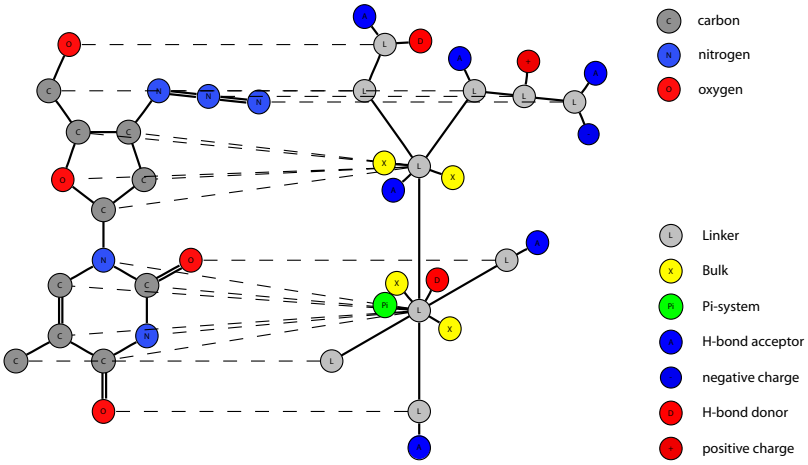


Figure 14.20: Zidovudine: Lewis structure and reduced graph.

14.8.6 Virtual Screening of Chemical Spaces

Investigator: Lars Kunert

Similarity based virtual screening methods are frequently used for analyzing molecular datasets. In order to obtain high performance – both in quality and run time – the underlying similarity function has to combine a high discriminative power with a fast evaluation procedure.

Maximum common substructures provide an intuitively reasonable view of the similarity between two molecules; but at present their computation is considered too time-consuming for use in practical similarity searches on larger data sets [1]. We have developed a method for indexing the chemical space of a set of molecules, such that a search for maximum common substructures can be done within a second.

The calculation of the this index is a preprocessing step that has to be done only once. As an example we indexed the drug-like subset of the ZINC database (two million molecules) [2] within a couple of days using a state-of-the-art workstation. The index can represent molecules as molecular graphs, hydrogen reduced graphs, or reduced graphs. Figure 14.20 shows the Lewis structure and the reduced graph of the HIV reverse transcriptase inhibitor zidovudine. The index has to contain the molecules which are used as queries later. A single molecule or a set of molecules can be used as a query. For a single query molecule, the result is a list of entries sorted by descending similarity. Each entry contains a similar molecule and the respective maximum common substructure. For a set of molecules, each list entry of the result is also tagged with the respective most similar query molecule. Weights for different types of substructures can be specified at query time. This allows for the use and fast training of different weighting schemes, e. g. for substructure searching or scaffold hopping.

Figure 14.21 shows the index – the maximum common subgraph DAG (mDAG) – of the set of colored graphs $\{A, B, C, D, E\}$. Every graph is canonically labeled (thereby it is ensured that only non-identical graphs are included into the mDAG), and for every pair of input

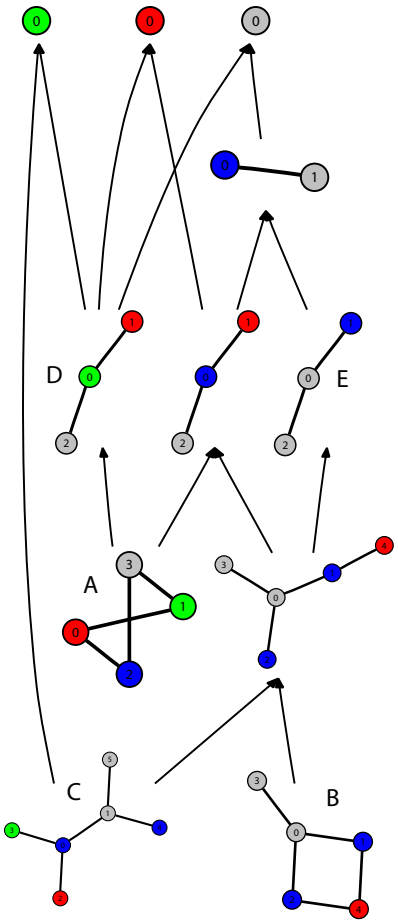


Figure 14.21: Maximum common subgraph DAG of the colored graphs A, B, C, D and E.

graphs and every positive weighting schema the DAG contains the corresponding maximum weight common subgraph.

The mDAG of a given set of molecules is computed in two steps: First the molecules are indexed according to their topology. The time complexity of this indexing step is linear in the output: The number of non-identical connected uncolored subgraphs contained in the input set. Like the mDAG the topological index is organized as a DAG, the tDAG of the input set. In a second step the tDAG is traversed level-by-level starting at the leaves and ending a the root. The algorithm traverses the complete space of colored connected subgraphs contained within the given input set. During the traversal all graphs that may become the maximum weight common subgraph of a pair of input graphs are identified by an approximate set-cover algorithm and included into the mDAG. The implementation of all mentioned algorithms is completed. At present we are working on the visualization of the results and on the optimization of the weighting schema.

At present the index is limited to a single representation per molecule. This can be generalized to multiple representations, for example to handle different protonation states. The presently used maximum common subgraph based similarity function can be extended to edit-distances or general topological transformations. This will allow for the combination of different representations within the same index. Using a combined index, the maximum common substructure of two molecules can consist of a single cyclic system represented as a reduced graph while the rest of the molecules is matched within the more specific hydrogen reduced graph representation.

References

[1] E. J. Barker, D. Buttar, D. A. Cosgrove, E. J. Gardiner, P. Kitts, P. Willett, and V. J. Gillet. Scaffold hopping using clique detection applied to reduced graphs. *J Chem Inf Comput Sci*, 46:503–511, 2006.

[2] J. J. Irwin and B. K. Shoichet. ZINC - a free database of commercially available compounds for virtual screening. *J Chem Inf Model*, 45:177–182, 2005.

14.8.7 Computational Approaches in Supramolecular Chemistry

Investigators: Andreas Steffen and Andreas Kämper

Supramolecular chemistry is defined as the chemistry of noncovalent interactions.[2] In most cases supramolecular systems consist of a host molecule and one or more guest molecules. Host molecules are synthesized molecules, which are commonly larger than guest molecules and possess a sizable central hole or cavity. In the context of this text we reference the host molecule as the synthetic receptor. The guest molecule may be a mono-atomic ion, a small molecular fragment or a molecule such as a drug that fits into this cavity and is complexed by the host molecule. The complex is held together by reversible noncovalent interactions, for example hydrogen bonds, van der Waals attractive forces, pi interactions or hydrophobic effects. The formation of the host-guest complex requires a complementarity of both complex partners with respect to interactions and steric arrangement and results in a unique structural relationship. In our department we develop computational tools that help chemists to rationally design novel host-guest complexes.

FlexR a novel docking tool for structure prediction of synthetic host-guest complexes

A significant prerequisite for the rational design of novel host-guest complexes is the existence of methods that reliably predict the structure of synthetic host-guest complexes. This is one of the computationally most challenging problems in supramolecular chemistry. Our idea was to transfer the concept of protein-ligand docking, which focuses on the structure prediction of complexes between ligands and protein binding sites, to synthetic host-guest complexes. Whereas current state-of-the-art docking tools all are able to handle the flexibility of the ligand, the efficient and reliable modeling of the protein's flexibility still remains a challenging task. In most attempts the assumption is made that the protein stays rigid during the binding process. Only some docking tools allow for limited flexibility, e.g. side-chain movement, movement of hydroxyl groups, or use ensembles of different protein conformations. Despite the fairly rough assumption of a rigid receptor, numerous success stories in which docking tools have led to novel drug candidates have been published. However, for a number of synthetic receptors the assumption of a rigid receptor cannot be carried over, as they can exhibit a high degree of flexibility, similar to ligands. In many cases not only the ligand but also the receptor adapts its conformation during complex formation. Thus, a docking tool for synthetic host-guest complexes has to tackle the problem of flexibility for both, the synthetic receptor and the guest molecules. The successful protein-ligand docking tool FlexX was developed in Thomas Lengauer's group at the GMD [3]. In our work we can rely on its basic algorithms and data structures and tailor and extend them for the purpose of our project. We have developed two related but considerably different algorithmical approaches for the structure prediction of hydrogen bond based synthetic host-guest complexes. The first paper [1] described our original approach, which was to perform a full conformational analysis of one of the two molecules and then to sequentially dock the other molecule into all generated conformations by means of the standard FlexX algorithm. Here, two different docking strategies were applied, forward and inverse docking. Whereas in forward docking the ligand is flexibly docked into a comprehensive sample of rigid receptor conformations, in inverse docking the roles of the two molecules are exchanged, i.e., the synthetic receptor is constructed around the ligand. This approach was successfully validated on a set of experimentally determined complex structures. Nevertheless a limitation of the method became apparent for cases in which both molecules of the complex exhibited a high degree of flexibility. We developed a new strategy to tackle this problem and implemented the two-sided incremental construction algorithm [4]. In contrast to the first approach here the structures of both molecules guest molecule and synthetic receptor are built up incrementally. Since the conformations of both molecules are unknown initially, we cannot use the conformation of one molecule to direct the generation of the conformation of the other molecule. Due to this fact a precomputation phase is introduced that determines putative interaction patterns between the two interacting molecules. This is done by means of distance estimations between interaction centers, the generation of a docking graph and determination of possible interaction patterns (cliques) within this graph by means of a clique search based algorithm. These interaction patterns direct the subsequent phase of complex construction. In comparison to the first approach a significant acceleration has been achieved for complexes that consist of highly flexible molecules, while the quality of the results has been maintained. The high efficiency of our two-sided incremental build-up approach originates from the fact that the

conformational space of both molecules is searched with respect to the counter molecule and thus the search space is significantly reduced. Together with some future modifications our tool will open new scenarios for the computer-assisted design of novel synthetic receptor-ligand complexes. In fact, even in its present state the speed of the approach allows for large-scale computation such as virtual screenings for an optimal ligand for a given synthetic receptor and vice-versa.

Virtual Screening for a Synthetic Receptor of Camptothecin

In a joint project together with researches from Saarland University and Ludwig-Maximilians-Universität Munich we developed a protocol for the design of a synthetic receptor based drug carrier system for camptothecin. Camptothecin and its derivatives represent a promising class of antineoplastic agents that exhibit a broad spectrum of activity against several types of cancer including colorectal and ovarian cancer. On the molecular level this class of drugs inhibits topoisomerase I which is a nuclear enzyme involved in the relaxation of DNA during cell replication and transcription. Unfortunately, its high therapeutic potential is hampered by poor solubility in water and low stability. Several attempts have been made to circumvent these difficulties by means of pharmaceutical formulations and inclusion in cyclodextrins. Our approach is based on an inverse virtual screening protocol where we first generated a virtual set of candidate receptors all of which are derivatives of beta-cyclodextrin, and then applied docking tools to generate complexes of camptothecin bound to each of the candidate receptors independently. Empirical scoring functions were used to estimate the binding free energy of all generated complexes and thus rank the candidate receptors. From the 10% top-ranking candidates we selected 9 promising for synthesis and experimental verification. Out of these 9 beta-cyclodextrin derivatives 6 exhibited a binding affinity to camptothecin significantly superior to any other known from literature. Our work is a successful demonstration of a computer-assisted design of novel synthetic receptors for a given guest molecule and opens up possibilities for the tailored design of drug delivery systems based on synthetic receptors.

Virtual Screening for new guest molecules of beta-cyclodextrin

Another way to optimize a given host-guest complex with respect to the binding free energy is to modify the guest molecule. In this study we report a combination of a ligand based virtual screening technique with a support vector machine (SVM) based regression model for predicting binding free energies of host-guest complexes. As an example we selected six known beta-cyclodextrin based host-guest complexes. The six guest molecules of beta-cyclodextrin served as query molecules for the ligand based virtual screening. A chemical compound dataset comprising 117,695 molecular entries served as the screening set. For all six query compounds a virtual screening was performed by means of a similarity based technique. We derived ranking lists out of the similarity scores. The best ranking 150 molecules of each of the ranking lists were then scored by means of the SVM regression model. The best scoring and most promising molecules of the six screening runs that were commercially available were selected for subsequent experimental verification. Altogether 16 compounds were purchased and their binding free energy to beta-cyclodextrin was determined by isothermal

microcalorimetric titration. Nine molecules exhibited a strong binding free energy. Five of these molecules even had a higher binding affinity than their corresponding query structures. The results are promising and introduce a well established concept from drug design to the field of supramolecular chemistry. This project has been accomplished in cooperation with researchers from Saarland University and Ludwig-Maximilian University Munich.

References

- [1] A. Kämper, J. Apostolakis, M. Rarey, C. M. Marian, and T. Lengauer. Fully automated flexible docking of ligands into flexible synthetic receptors using forward and inverse docking strategies. *Journal of Chemical Information and Modeling*, 46(2):903–911, March 2006.

[2] J. Lehn. Supramolecular chemistry - scope and perspectives molecules, supermolecules, and molecular devices. *Angew Chem Int Ed*, 27(1):89–112, Jan 1988.

[3] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–489, Aug 1996.

- [4] A. Steffen, A. Kämper, and T. Lengauer. Flexible docking of ligands into synthetic receptors using a two-sided incremental construction algorithm. *Journal of Chemical Information and Modeling*, 46(4):1695–1703, July 2006.

14.9 System Administration

System administration covers maintaining the hard- and basic software (Subsection 14.9.1) for individual group members, providing the infrastructure for scientific services(Subsection 14.9.2) as well as rendering methods available as web services (Subsection 14.9.3).

14.9.1 Hard- and Software Configuration

Investigator: Joachim Büch

Software Packages and Bioinformatics Software

Each group member has a Linux desktop or a notebook with dual boot Windows and Linux for his or her personal use. Some are using both desktop and notebook. Due to the diverse research interests in the group a lot of complex software such as bioinformatics toolboxes for scripting languages (Perl, Python, PHP) and statistical software packages (R, Matlab), multiple alignment tools and software for molecular modelling and annotation servers must be available for the scientific research projects.

To keep all this software up to date on 70 desktops and notebooks, 100 compute cluster nodes and 4 application servers, packages are provided, which can be remotely installed by a mechanism that was developed in the central computer support group of the institute.

The software has to be provided not only for 32-bit computer systems, but also for servers with 64-bit architecture, because calculations with huge data amounts need significantly more than 4 GB of main memory.

Compute Clusters

There are three compute clusters available for the group:

- The institute's Sun Fire 15000 with 58 UltraSPARC III processors and 160GB main memory is still used for processes that need a large amount of memory. Bioinformatics software had therefore to be compiled for Sun Solaris.
- The group runs a Dell compute cluster with 30 Dell's Power Edge 2650 server nodes with Dual Xeon 3.1 GHz processors and 4 GB RAM per node. The installed Message Passing Interface allows for running parallel programs on the nodes. Projects working with Gromacs software and DynaCell, which was developed in the group, are the primary users of this cluster.
- The institute provides a Linux compute cluster with 100 nodes, each of whose nodes is a Suns V20z with the 64-bit architecture of two dual core Opterons with 8 GB memory. Job queuing is done by Sun's "N1 Grid Engine" software.

Database Servers

Since most of the projects use MySQL we have 3 database servers running MySQL. Version 4.1 is running on a Sun Ultra 80 with 4 UltraSPARC II processors, 4 GB of main memory and 600 GB disk space. Another database that serves as backend system for the Internet content is hosted on a Dual Xenon machine with 270 GB disk space available. Newer projects benefit from new features of MySQL 5.0. This 64-bit version is installed together with Oracle on a new Sun V40z server with 16 GB memory. The Oracle database was migrated to Oracle 10g Release 2, because the new research project Epigenetics (see Subsection 14.7.4) needs the enhanced features of XML DB. Disk space of 1.2TB for Oracle and MySQL data on this server are provided by a Storage Area Network (SAN).

Web Servers

Some of our websites still reside on our old Sun Fire 280R with 2 UltraSPARC III processors and Solaris 9, because of problems, when upgrading to Solaris 10. The majority of the web services is hosted now on a new Sun V40z web server with two 2.5MHz Opteron processors. They had to be suited for this new environment.

Storage Area Network (SAN)

The group purchased and installed a new disk array that was attached over a Fiber channel switch to the above described database, application and web servers, providing 3.6 TB of disk space.

Future Hardware Plans

Another disk array is just being installed, which allows for mirroring the existing one. Since our Oracle and MySQL databases are growing rapidly, we plan to acquire another 3.6TB of disks for our SAN. Since the Oracle installation under Debian Linux only runs in 32-bit

mode, we need another database server, which allows to run 64-bit Oracle software and provides enough memory to the database instance to get more performance.

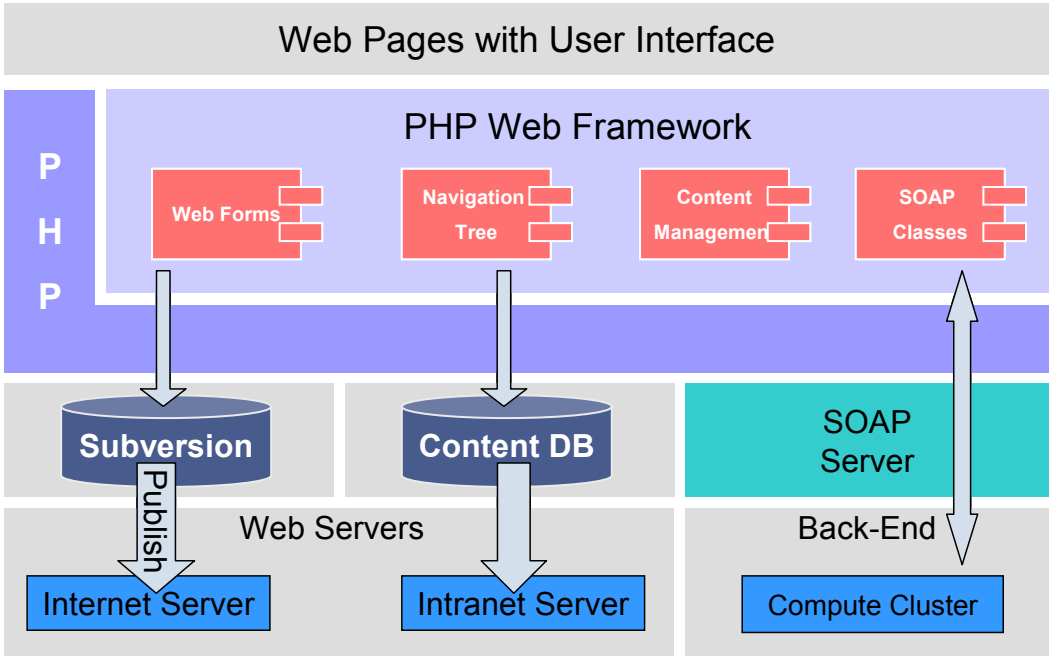


Figure 14.22: The web framework automates the creation of new websites.

14.9.2 Infrastructure

Investigator: Joachim Büch

PHP Web Framework and Intranet Corporate Design

Publicly available content management systems like Typo3 are not flexible enough to fulfill the needs of our web services, because each project uses different programming libraries providing the best tools for solving the particular problem. We searched for a technique allowing for separating the web front-end from the program logic. Additionally a basic template for the websites should automatically provide the MPI's corporate identity. Moreover the websites should be independent of the web servers hardware and operating system and the base URL of the project, allowing for running the same web pages on different web servers. We therefore developed an object-oriented PHP class library taking all configuration data and basic information for the website out of a database (see Figure 14.22). This framework also provides classes for web forms. Web sites can be developed and tested on an internal web server and the revision control system Subversion is used to keep track of changes in the web pages.

Intranet

In our intranet we provide information about installed software and hardware, seminar talks and so on. Some of the pages using our web framework others use an installed Wiki. Additionally a helpdesk systems helps to coordinate installation needs of the group with the system administrator.

Databases

Some bioinformatics databases come as flat files, but the majority of projects now use MySQL or Oracle databases for performance reasons. Specifically the project on Epigenetics (see Subsection 14.7.4), which currently stores 250GB 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 has to be done to reach the best performance.

For the GENO2PHENO project (see Subsection 14.4.1), we are hosting a copy of the *Arevir* database [2] 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.

Consulting of Group Members

Supporting the group members, when designing databases and web services for new projects is as important as introducing new master students into the hard- and software environment of the institute. Consulting therefore plays a major role in system administration.

14.9.3 Web Services

Investigator: Joachim Büch

Bioinformatics projects

Large data models for Protein Structure and Function Predictions (see Section 14.5), Analysis of Protein Binding sites (see Subsection 14.5.3), HIV Drug Resistance prediction ,HIV therapy optimization (see Subsection 14.4.1) and the Epigenetics project (see Subsection 14.7.4) are stored in Oracle Databases. Online available web services, which may benefit from those databases, become more and more important, because it is not possible anymore to distribute stand-alone software tools together with data models. Therefore, in contrast to the other groups in the institute, the bioinformatics group offers a number of internet-based services to the bioinformatics community.

Typically, together with the researchers, the system administrator chooses the best fitting software concept for the needs of the individual projects. The system administrator is responsible for the maintenance of following web services listed in Table 14.1.

Genafor	http://www.genafor.org
Geno2Pheno (14.4.1)	http://www.geno2pheno.org
Coreceptor (14.4.2)	http://coreceptor.bioinf.mpi-sb.mpg.de
HBV Drug Resistance	http://www.genafor.org/hbv/hbvpredict.php
Arby [1]	http://arby.bioinf.mpi-inf.mpg.de
CalSpec (14.5.6)	http://bioinf.mpi-sb.mpg.de/projects/CalSpec
BiQ Analyzer (14.7.4)	http://biq-analyzer.bioinf.mpi-inf.mpg.de
topGO (14.7.1)	http://topgo.bioinf.mpi-sb.mpg.de
DomainNetworkBuilder (14.6)	http://med.bioinf.mpi-sb.mpg.de/domainnet
NetworkAnalyzer (14.6)	http://med.bioinf.mpi-sb.mpg.de/netanalyzer
MHC Peptide Interaction (14.8.3)	http://bioinf.mpi-sb.mpg.de/projects/mhc
NOXclass (14.5.4)	http://noxclass.bioinf.mpi-inf.mpg.de
GOTax (14.6.1)	http://gotax.bioinf.mpi-inf.mpg.de
Recco (14.4.4)	http://recco.bioinf.mpi-sb.mpg.de
ROCR [1]	http://rocr.bioinf.mpi-sb.mpg.de
STRuster (14.5.2)	http://struster.bioinf.mpi-sb.mpg.de
IRECS (14.8.2)	http://irecs.bioinf.mpi-sb.mpg.de
EURESIST	http://euresist.bioinf.mpi-sb.mpg.de
Mtreemix [1]	http://mtreemix.bioinf.mpi-sb.mpg.de
ViralDAS	http://viraldas.bioinf.mpi-sb.mpg.de

Table 14.1: Overview of the web services maintained by Department 3. Some web services were built for external cooperations, some were discussed in the previous biennial report [1], those followed by parentheses () are discussed in the corresponding sections of this report.

Calculations for web services with short response time run directly on the Internet server and the result is presented immediately in the browser window. Services that need 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 institute's 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.

The web interfaces for the web services are mostly developed with the PHP Framework described above. This allows for strict separation between the web design with the look and feel on the website and the program logic behind (MVC Model View Controller concept).

We have consolidated existing web services of the group. With consolidation we mean, rendering synchronous web services faster by replacing scripting languages with native program code in C++. Asynchronous web services must become more robust i.e. the complex data flow among the front-end, the database and the computation on the compute cluster has to be supervised and controlled in order to detect possible exceptions. The web services

provided by GENO2PHENO (see Subsection 14.4.2) are currently being redesigned as a first step in this direction (see Figure 14.23).

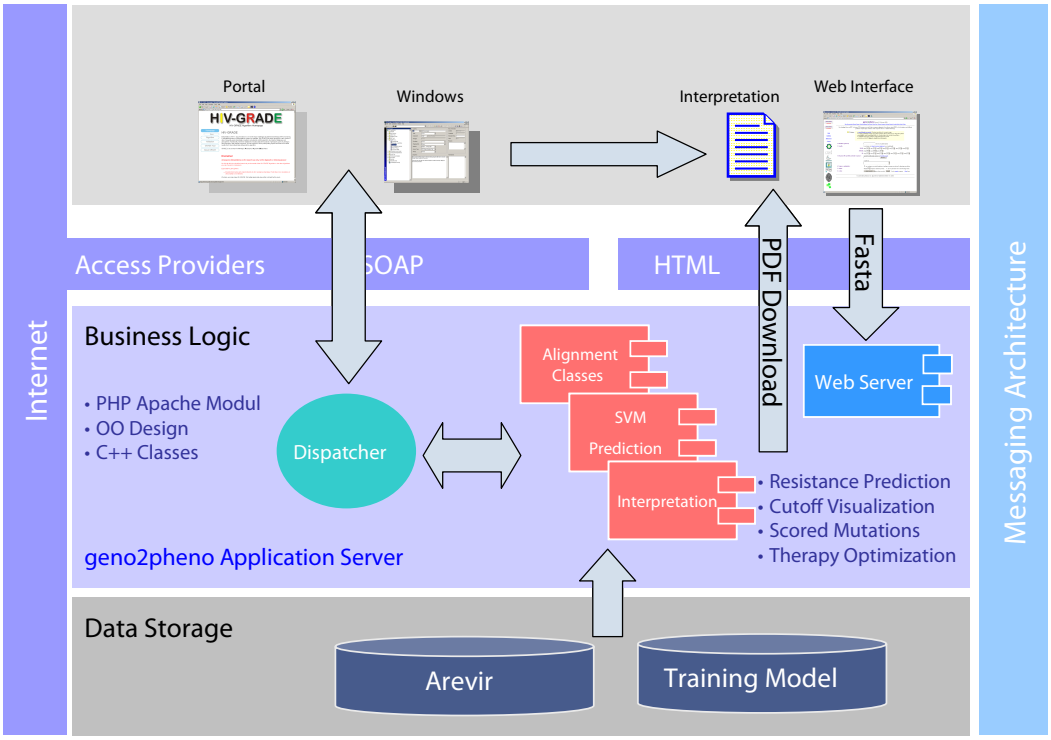


Figure 14.23: The new software design of GENO2PHENO uses the SOAP protocol to link the GUI with the server-side back-end, thus allowing to use different clients for GENO2PHENO. For the sake of a better performance, all program objects will be linked together in an Apache module.

Conferences

We are providing the website `http://www.genafor.org/events.php` for the annual Arevir meeting in Bonn.

In 2005, we helped organizing the "BioSapiens-viRgil Workshop on Bioinformatics for Viral Infections". The corresponding website `http://workshop2005.bioinf.mpi-sb.mpg.de` for the conference with online poster registration was also done with our PHP Framework, which was described above.

References

[1] M.-P.-I. für Informatik. Seventh biennial report. Seventh Biennial Report, 2005.

- [2] 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. In U. Leser, F. Naumann, and B. Eckman, eds., *Data Integration in the Life Sciences, Third International Workshop, DILS 2006*, July 2006, *LNCS 4075*, pp. 185–194. Springer.

14.10 Academic Activities

14.10.1 Journal Positions

Thomas Lengauer is on the editorial board of

- *Bioinformatics (Associate Editor)* (since 1996),
- *Comparative and Functional Genomics* (since 2003),
- *Discrete Applied Mathematics* (since 1989),
- *IEEE-ACM Transactions on Computational Biology and Bioinformatics* (since 2004),
- *Journal of the ACM (Area editor: Computational Biology)* (1991–2006),
- *Journal of Computational Biology* (since 1997),
- *Springer Lecture Notes in Bioinformatics* (since 2003).

14.10.2 Conference and Workshop Positions

Membership in Program Committees

Thomas Lengauer:

- *Fourth European Conference on Computational Biology (ECCB 2005)*, Madrid, Spain, September 2005.
- *Tenth Annual International Conference on Computational Molecular Biology (RECOMB 2006)*, Venice, Italy, April 2006.
- *Fifth European Conference on Computational Biology (ECCB 2006)*, Eilat, Israel, September 2006.
- *Eleventh Annual International Conference on Computational Molecular Biology (RECOMB 2007)*, Oakland, April 2007.

Mario Albrecht:

- *ISMB/ ECCB 2007, 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB).*

Adrian Alexa:

- *ISMB/ ECCB 2007, 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB).*

Francisco Domingues:

- *ISMB/ ECCB 2007, 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB).*

Jörg Rahnenführer:

- *GMDS 2006, 51. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*, Leipzig, September 2006.
- *ISMB/ ECCB 2007, 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB).*

Oliver Sander:

- *ISMB/ ECCB 2007, 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB).*

Ingolf Sommer:

- *ISMB/ ECCB 2007, 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB).*

Membership in Organizing Committees

Thomas Lengauer:

Conference/Workshop Organizations:

- *BioSapiens-viRgil Workshop on Bioinformatics for Viral Infections*, Bonn, Germany, 21.-23.09.2005 (with Daniel Hoffmann).
- *Workshop on Protein Bioinformatics*, Koç University, September 6-8, 2006, (with Burak Erman and Volkhard Helms).
- *Fifteenth International Conference on Intelligent Systems for Molecular Biology, jointly with the sixth European Conference on Computational Biology (ISMB/ECCB 2007)*, Vienna, July 21–25, 2007 (conference chair, with Burkhard Rost, Peter Schuster, Conference Co-chairs).

Jörg Rahnenführer:

- *Frankfurter Stochastiktage 2006: Organizer of Sektion 11: Statistik in Biowissenschaften und Medizin*, Frankfurt, March 2006.

14.10.3 Invited Talks and Tutorials

Thomas Lengauer:

Talks on HIV Bioinformatics

- Conf. on Therapeutic Application of Computational Biology, Hinxton, UK, September 2005
- MPG-CAS Institute for Computational Biology, Shanghai, October 2005.
- Koç University, Istanbul, Turkey, October 2005.
- BioScope-IT Kickoff Meeting, Leuven, November 2005.
- NRW Akademie der Wissenschaften, February 2006.
- MPG Workshop on Systems Biology, March 2006.
- CUBIC-CDFD Workshop, Univ. Köln, April 2006 "Informatik überzeugt", Univ. Paderborn, September 2006.
- "Wissenschaft für jedermann", Deutsches Museum München, October 2006.
- Workshop on Bioinformatics and Modeling in Biomedicine, Luxembourg, November 2006.
- Workshop on Innovative Technologies in the Preventive Health Sector, Enschede, Netherlands, November 2006.
- Bioinformatics Munich: From Genomes to Systems Biology, November 2006.
- Fachbereich Informatik, RWTH Aachen, November 2006.
- CMBI, Universität Nijmegen, January 2007.
- Workshop über Systembiologie der CPTS der MPG, Leipzig, January 2007.

Overview of Bioinformatics in Phamarceutical Research

- Cell Course, Saarland University, March 2006.

Perspectives of Computational Biology

- Perspektiven der Informatik, IBFI Schloss Dagstuhl, November 2006.

Mario Albrecht:

- *Analysis of disease-associated protein interaction networks*, Invited talk, Annual NGFN Project Meeting for SMP-Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany, October 2006.
- *Functional evaluation of human protein-protein networks*, Workshop on Protein Bioinformatics, Koç University, Istanbul, Turkey, September 2006.
- *Functional evaluation of predicted human protein networks*, 2nd CSHL Meeting on Interactome Networks, Hinxton, United Kingdom, August 2006.
- *Protein Bioinformatics Support*, Annual NGFN Project Meeting for Diseases due to Environmental Factors, Charité University Medicine, Berlin, Germany, July 2006.
- *Bioinformatics analysis of disease-associated proteins*, Annual NGFN Project Meeting for SMP-Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany, December 2005.

– *Decomposing protein networks into domain-domain interactions*, 4th European Conference on Computational Biology (ECCB), Madrid, Spain, September 2005.

Iris Antes:

– *Inhibitor design based on homology modeled protein structure*, Invited talk, ChemBioNet Conference, Frankfurt am Main, Germany, December 2005.
– *Docking into flexible binding sites: ATPase, CYP11B2, and P2OX*, Invited talk, Biotechnology Seminar, Saarland University, Saarbrücken, Germany, July 2006.
– *DynaDock: Inhibitor based refinement of homology modeled protein structures for molecular docking*, Invited talk, 3DSig Meeting, Fortaleza, Brasil, August 2006.
– *DynaPred: A structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations*, Conference talk, 14th Annual International Conference on Intelligent Systems in Molecular Biology (ISMB), Fortaleza, Brasil, August 2006.
– *DynaPred: A structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations*, Conference talk, German Conference on Bioinformatics 2006, Tübingen, Germany, September 2006.
– *Joining bioinformatics and biophysical methods for drug design*, Invited talk, Biochemistry Seminar, University of Zürich, Zürich, Switzerland, October 2006.
– *Combined bioinformatics and biophysical methods for molecular docking and protein modeling*, Invited talk, Bio-Mathematics and Informatics Symposium, CWI Amsterdam, April 2007.

Jörg Rahnenführer:

– *Genetic tumor progression scores*, Invited talk, IMISE, Leipzig, July 2005.
– *Exploratory data analysis for microarrays*, NGFN – Courses in Practical DNA Microarray Analysis, Saarbrücken, September 2005.
– *Design und Analyse von Microarrays: Wie man mit Statistik nicht in der Datenflut ertrinkt*, Invited talk, Kongress Deutsche Gesellschaft für Urologie, Düsseldorf, September 2005.
– *Genetische Tumorprogressions-Scores*, Invited talk, Institut für Informatik, Düsseldorf, February 2006.
– *Genetische Tumorprogressions-Scores*, Invited talk, IMIS, Kiel, March 2006.
– *Genetic tumor progression scores*, Frankfurter Stochastik-Tage 2006, Frankfurt, March 2006.
– *Genetische Tumorprogressions-Scores*, Invited talk, Institut für funktionelle Genomik, Regensburg, April 2006.
– *Statistische Methoden für die biologische Aufklärung von Krankheitsverläufen aus genomischen Daten*, Invited talk, Medizinische Fakultät, Regensburg, April 2006.
– *Exploratory data analysis for microarrays*, NGFN – Courses in Practical DNA Microarray Analysis, München, May 2006.

- *From a gene list to biological function – Scoring Gene Ontology terms*, NGFN – Courses in Practical DNA Microarray Analysis, München, May 2006.
- *Praktische Datenanalyse für Microarrays*, Tutorial (with Rainer Spang) at GMDS 2006, Leipzig, September 2006.
- *Exploratory data analysis for microarrays*, NGFN – Courses in Practical DNA Microarray Analysis, Heidelberg, November 2006.
- *Exploratory data analysis for microarrays*, NGFN – Courses in Practical DNA Microarray Analysis, Heidelberg, March 2007.
- *Statistical analysis of genetic changes in tumor cells for estimating cancer progression*, DAGStat Tagung 2007: Statistik unter einem Dach, Bielefeld, March 2007.

Ingolf Sommer:

- *Prediction and Classification of Protein Structures and Complexes*, Workshop on Protein Bioinformatics, Koç University, Istanbul, Turkey, September 2006.
- *From Protein Structure to Function*, Invited talk, Bio-Mathematics and Informatics Symposium, CWI Amsterdam, April 2007.

14.11 Teaching Activities

Summer Semester 2005

Courses:

Computational Biology / Bioinformatics II (Lecturer: Thomas Lengauer, Tutor: Christoph Bock)
Computer-Aided Drug Design (Lecturer: Andreas Kämper, Tutor: Andreas Kämper)
The Elements Of Statistical Learning II (Lecturer: Jörg Rahnenführer, Tutor: Adrian Alexa)

Winter Semester 2005/2006

Courses:

The Elements Of Statistical Learning I (Lecturer: Thomas Lengauer, Tutor: Adrian Alexa)

Summer Semester 2006

The Elements Of Statistical Learning II (Lecturer: Jörg Rahnenführer, Tutor: Laura Tolosi)
Computer-Aided Drug Design (Lecturer: Andreas Kämper, Tutor: Andreas Steffen)

Winter Semester 2006/2007

Courses:

The Elements Of Statistical Learning I (Lecturer: Thomas Lengauer, Tutor: Laura Tolosi)

Master's Theses

- Adrian Alexa: Integrating the GO graph structure in scoring the significance of Gene Ontology terms, 2005
- Andreas Schlicker: A global approach to comparative genomics: Comparison of functional annotation over the taxonomic tree, 2005
- Weng-In Siu: Computational prediction of MHC-peptide interactions, 2005.
- Yassen Assenov: Topological Analysis of Biological Networks, 2006
- Thomas Binsl: Feature Prints: A new Method for Calculating Molecular Similarities, 2006
- Xin Guo: Improving Fold Recognition using Protein and Domain Interactions, 2006
- Konstantin Halachev: EpiGRAPH-regression: A toolkit for (epi-)genomic correlation analysis and prediction of quantitative attributes, 2006
- Carola Huthmacher: Decomposing Protein Networks Into Domain-Domain Interactions, 2006
- Stefan Immesberger: Heuristics to Speed up Profile-Profile Alignment, 2006
- Melanie Kaspar: Automated Conformational Analysis of Macrocyclic Ring Systems, 2006
- Oliver Mueller: Using Shape Retrieval Techniques for Identifying Similar Protein Binding Sites, 2006
- Laura Tolosi: Analysis of Array CGH Data for the Estimation of Genetic Tumor Progression, 2006
- Nils Weinhold: Inference of Protein Function based on Functionally Conserved Regions in Sequence and Structure Space, 2006
- Jasmina Bogojeska: Stability Analysis of Oncogenetic Trees Mixture Models, 2007

14.12 Dissertations, Habilitations, Offers, Awards

14.12.1 Dissertations

Dissertations supervised by Thomas Lengauer:

- Mario Albrecht: *Combining protein structure prediction with experiments and functional information*, Dr. rer. nat. (Computer Science), Saarland University, Saarbrücken, June 2006.

14.12.2 Habilitations

Jörg Rahnenführer

- *Statistical methods for the biological interpretation of genome-wide measurements*, PD (Bioinformatics), Saarland University, Saarbrücken, October 2006.

14.12.3 Offers for Faculty Positions

Jörg Rahnenführer

- Universität Kiel, 2006.
- Universität Dortmund, 2006.

14.12.4 Awards

- Bernd Wullich et al. *Entwicklung genetischer Progressionsscores als prädiktiveMarker zur Risikobeurteilung von Prostatakarzinomen*, **Wolfgang-Hepp-Prize**, Düsseldorf, September 2005.
- Oliver Mueller et al., *Using Shape Retrieval Techniques for Identifying Similar Protein Binding Sites*, **Best Poster Award, German Conference on Bioinformatics (GCB)**, Tuebingen, October 2006.
- Christoph Hartmann et al., *IRECS: Prediction of Side-Chain Conformation Ensembles*, **Best Poster Award, Drug Discovery Workshop**, Marburg, March 2007.
- Oliver Sander et al. *Prediction of HIV-1 Coreceptor Usage Based on Structural Descriptors of the gp120 V3 Loop*, **Best Poster Presentation Award, 5th European HIV Drug Resistance Workshop**, Cascais, Portugal, March 2007.

14.13 Grants and Cooperations

There are a number of projects funded with third-party money that are described in the succeeding subsections.

In addition, we have several joint projects with partners on the Saarbrücken and Homburg campuses that are in a introductory stage and have not gained outside funding yet. These include the project on HIV evolution with Prof. Meyerhans (Virology, Homburg), see also Section 14.4.4, inhibitor design for aldosterone synthase with Prof. Hartmann (Pharmaceutical Chemistry, Saarbrücken) 14.8.5 and geometric aspects of the analysis of protein binding-sites with Prof. Weickert (Mathematics and Computer Science, Saarbrücken) 14.5.3.

We are in a constant dialog with the GMD startup BioSolveIT GmbH, Sankt Augustin.

Coordinating Activities

DFG Projects

Metabolomics (partially funded through CBI)

In this project, our experimental partner analyzes yeast-knockouts that are being fed with 13C-labeled glucose. Then the cells are submitted to labeled amino-acid screening over a certain time interval. We have contributed to the software for analyzing the involved mass spectra and selecting the relevant knock-outs. We develop bioinformatics methods to analyze the primary screening data (labeled amino-acid abundances) and the resulting metabolite fluxes calculated with state-of-the-art flux determination software (see Section 14.5.6).

Partners:

- Max Planck Institute for Informatics (Prof. Thomas Lengauer, Priti Talwar)
- Biotechnology, CBI (Prof. Elmar Heinzle, Vidya Mangadu, Dr. Christoph Wittmann)

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 14.8.5).

Partners:

- Max Planck Institute for Informatics (Prof. Thomas Lengauer, Dr. Iris Antes)
- Saarland University, Pharmacology (Prof. Rolf W. Hartmann, Ursula Müller, Sarah Ulmschneider)

Design of Artificial Receptors

In this project, we are developing methods for the *de novo* design of artificial (synthetic) receptors. We have developed the new tool FlexR, which uses the docking technique of FlexX for structure prediction of complexes between these artificial receptors and their ligands, taking flexibility of both partners into account (see Section 14.8.7).

Partners:

- Max Planck Institute for Informatics (Dr. Andreas Kämper)
- Düsseldorf University (Prof. Christel M. Marian)
- University of Munich (Dr. Joannis Apostolakis)

Sequence Analysis for HCV

This project is part of a Clinical Research Group of DFG on analyzing the Hepatitis C Virus with experimental and bioinformatics methods. The Research Group is coordinated by Prof. Stefan Zeuzem from the Medical Faculty of University of the Saarland (Gastroenterology). In the project we are analyzing the correlations between the genotypic variants of HCV and their phenotypic resistance behavior (see Section 14.6.3).

Partners:

- DFG Clinical Research Group on Hepatitis C

Structure Analysis for HCV

This project is part of a Clinical Research Group of DFG on analyzing the Hepatitis C Virus with experimental and bioinformatics methods. The Research Group is coordinated by Prof. Stefan Zeuzem from the Medical Faculty of University of the Saarland (Gastroenterology). In the project we are modeling the structures of important HCV proteins (see Section 14.6.3).

Partners:

- DFG Clinical Research Group on Hepatitis C

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/`) aims to improve the bioinformatics and statistics support for the design and analysis of gene expression data in the NGFN. Basic techniques are taught in regularly held courses on the analysis of gene expression data. In addition, scientific contributions by the MPI group focus on method development for an enhanced biological interpretation of expression data (see Section 14.7.1).

Partners:

- Members of the NGFN SMP (Systematic Methodological Platform) Bioinformatics, directed by Prof. Roland Eils, DKFZ Heidelberg
- Max Planck Institute for Informatics (Prof. Thomas Lengauer, Dr. Jörg Rahnenführer)

Bioinformatics Services for Environmental Diseases

This project is funded by the NGFN (Nationales Genomforschungsnetzwerk). Our task here is to provide structural and functional annotations for proteins that are targeted experimentally within the Genomic Network on Environmental Diseases, see also Section 14.6.3).

Partners:

- Members of the NGFN Genomic Network on Environmental Diseases, directed by Prof. Stefan Schreiber, University of Kiel
- Max Planck Institute for Informatics (Prof. Thomas Lengauer, Mario Albrecht)

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 the workpackages 9 (on inferring protein function from sequence and structure) and 15 (on bioinformatics for infectious diseases).

Partners:

- Members of the Biosapiens Network

Eu*Resist*

The Eu*Resist* project aims at developing a European integrated system for clinical management of antiretroviral drug resistance. The system will provide 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 huge European integrated data set will be created, linking some of the largest

existing resistance databases. In this project, we are especially engaged in the workpackages 3 (prediction models) and 4 (comparison and combination of models).

Partners:

- Informa S.r.l. (Francesca Incardorna)
- University of Siena (Prof. Maurizio Zazzi)
- Karolinska Institute (Prof. Anders Sonnerborg)
- 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)

14.14 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] M. Albrecht. LRRK2 mutations and parkinsonism. *The Lancet*, 365(9466):1230–1230, 2005.

[2] M. Albrecht, D. Choubey, and T. Lengauer. The HIN domain of IFI-200 proteins consists of two OB folds. *Biochemical and Biophysical Research Communications*, 327(3):679–687, 2005.

[3] M. Albrecht, C. Huthmacher, S. C. Tosatto, and T. Lengauer. Decomposing protein networks into domain-domain interactions. *Bioinformatics*, 21(Suppl. 2):ii220–ii221, 2005.

[4] M. Albrecht and F. L. W. Takken. Update on the domain architectures of NLRs and R proteins. *Biochemical and Biophysical Research Communications*, 339(2):459–462, 2006.

[5] 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.

[6] 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.

[7] I. Antes, C. Merkwirth, and T. Lengauer. POEM: Parameter optimization using ensemble methods: Application to target specific scoring functions. *Journal of Chemical Information and Modeling*, 45(5):1291–1302, 2005.

[8] I. Antes, 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.

[9] M. Balduin, N. Beerenwinkel, S. Sierra, M. Däumer, J. Rockstroh, M. Oette, G. Fätkenheuer, B. Kupfer, D. Hoffmann, J. Selbig, H. Pfister, and R. Kaiser. Evolution of HIV resistance during treatment interruption in experienced patients and after restarting a new therapy. *J Clin Virol*, 34(4):277–287, 2005.

[10] N. Beerenwinkel, M. Däumer, T. Sing, J. Rahnenführer, T. Lengauer, J. Selbig, D. Hoffmann, and R. Kaiser. Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *The Journal of Infectious Diseases*, 191(11):1953–1960, 2005.

[11] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer. Learning multiple evolutionary pathways from cross-sectional data. *Journal of Computational Biology*, 12(6):584–598, 2005.

[12] N. Beerenwinkel, J. Rahnenführer, R. Kaiser, D. Hoffmann, J. Selbig, and T. Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107, 2005.

[13] N. Beerenwinkel, T. Sing, T. Lengauer, J. Rahnenführer, K. Roomp, I. Savenkov, R. Fischer, D. Hoffmann, J. Selbig, K. Korn, H. Walter, T. Berg, P. Braun, G. Fätkenheuer, M. Oette, J. Rockstroh, B. Kupfer, R. Kaiser, and M. Däumer. Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics*, 21(21):3943–3950, 2005.

[14] C. Bock. Bioinformatik: Neue Strategien gegen Krebs. *Deutsches Ärzteblatt PRAXiS*, 103(36):10–11, 2006.

[15] 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.

[16] 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.

[17] A. Boeddrich, S. Gaumer, A. Haacke, N. Tzvetkov, M. Albrecht, B. O. Evert, E. C. Müller, R. Lurz, P. Breuer, N. Schugardt, S. Plaßmann, K. Xu, J. M. Warrick, J. Suopanki, U. Wüllner, R. Frank, U. F. Hartl, N. M. Bonini, and E. E. Wanker. An arginine/lysine-rich motif is crucial for VCP/p97-mediated modulation of ataxin-3 fibrillogenesis. *EMBO Journal*, 25(7):1547–1558, 2006.

[18] J. Bojunga, C. Welsch, I. Antes, M. Albrecht, T. Lengauer, and S. Zeuzem. Structural and functional analysis of a novel mutation of CYP21B in a heterozygote carrier of 21-hydroxylase deficiency. *Human Genetics*, 117(6):558–564, 2005.

[19] S. Castellano, A. V. Lobanov, C. Chapple, S. V. Novoselov, M. Albrecht, D. Hua, A. Lescure, T. Lengauer, A. Krol, V. N. Gladyshev, and R. Guigó. Diversity and functional plasticity of eukaryotic selenoproteins: Identification and characterization of the SelJ family. *Proceedings of the National Academy of Sciences*, 102(45):16188–16193, 2005.

[20] C. M. Costello, N. Mah, R. Häsler, P. Rosenstiel, G. H. Waetzig, A. Hahn, T. Lu, Y. Gurbuz, S. Nikolaus, M. Albrecht, J. Hampe, R. Lucius, G. Klöppel, H. Eickhoff, H. Lehrach, T. Lengauer, and S. Schreiber. Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays identifies novel candidate disease genes. *PLoS Medicine*, 2(8):771–787, 2005.

[21] A. V. Das, J. James, J. Rahnenführer, W. B. Thoresona, S. Bhattacharyaa, X. Zhao, and I. Ahmad. Retinal properties and potential of the adult mammalian ciliary epithelium stem cells. *Vision Research*, 45(13):1653–1666, 2005.

[22] 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.

[23] W. Gu, M. Kofler, I. Antes, C. Freund, and V. Helms. Alternative binding modes of polyproline peptides binding to the GYF domain. *Biochemistry*, 44(17):6404–6415, April 2005.

[24] A. Hahn, J. Rahnenführer, P. Talwar, and T. Lengauer. Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics*, 6(112):1–11, 2005.

[25] 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.

[26] C. Hartmann, I. Antes, and T. Lengauer. IRECS: A new algorithm of the selection of most probable ensembles of side-chain conformations in protein models. *Protein Science*, 15, 2007.

[27] D. M. Heckmann-Pohl, S. Bastian, S. Altmeier, and I. Antes. Improvement of the fungal enzyme pyranose 2-oxidase using protein engineering. *Journal of Biotechnology*, 124(1):26–40, 2006.

[28] B. Heil, J. Ludwig, H. Lichtenberg-Fraté, and T. Lengauer. Computational recognition of potassium channel sequences. *Bioinformatics*, 13(22):1562–1568, 2006.

[29] G. Hessler, M. Zimmermann, H. Matter, A. Evers, T. Naumann, T. Lengauer, and M. Rarey. Multiple-ligand-based virtual screening: Methods and applications of the MTree approach. *Journal of Medicinal Chemistry*, 48(21):6575–6584, 2005.

[30] A. Kämper, J. Apostolakis, M. Rarey, C. M. Marian, and T. Lengauer. Fully automated flexible docking of ligands into flexible synthetic receptors using forward and inverse docking strategies. *Journal of Chemical Information and Modeling*, 46(2):903–911, March 2006.

[31] 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.

[32] 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*, p. Published online January 17, 2007.

[33] C. Lehmann, M. Däumer, I. Bousaad, T. Sing, N. Beerenwinkel, T. Lengauer, N. Schmeisser, C. Wyen, G. Fätkenheuer, and R. Kaiser. Stable coreceptor usage of HIV in patients with ongoing treatment failure on HAART. *Journal of Clinical Virology*, 37(4):300–304, 2006.

[34] T. Lengauer. Informatics (computational biology). In R. A. Meyers, ed., *Growth Factors and Oncogenes in Gastrointestinal Cancers to Informatics (Computational Biology)*, *Encyclopedia of Molecular Cell Biology and Molecular Medicine*, vol. 6, pp. 567–626. Wiley-VCH, Weinheim, Germany, 2005.

[35] T. Lengauer. Bioinformatics - from genomes to therapies - 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.

[36] 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.

[37] T. Lengauer and C. Merkwirth. Automatic generation of complementary descriptors with molecular graph networks. *Journal of Chemical Information and Modeling*, 45(5):1159–1168, 2005.

[38] T. Lengauer and T. Sing. Bioinformatics-assisted anti-HIV therapy. *Nature Reviews Microbiology*, 4:790–797, October 2006.

[39] J. Maydt and T. Lengauer. Recco: recombination analysis using cost optimization. *Bioinformatics*, 22(9):1064–1071, February 2006.

[40] U. Mihm, N. Grigorian, C. Welsch, E. Herrmann, B. Kronenberger, G. Teuber, M. von Wagner, W.-P. Hofmann, M. Albrecht, T. Lengauer, S. Zeuzem, and C. Sarrazin. Amino acid variations in hepatitis C virus p7 and sensitivity to antiviral combination therapy with amantadine in chronic hepatitis C. *Antiviral Therapy*, 11(4):507–519, 2006.

[41] D. Neumann, O. Kohlbacher, C. Merkwirth, and T. Lengauer. A fully computational model for predicting percutaneous drug absorption. *Journal of Chemical Information and Modeling*, 1(46):424–429, 2006.

[42] G. Plotz, C. Welsch, L. Giron-Monzon, P. Friedhoff, M. Albrecht, A. Piiper, R. M. Biondi, T. Lengauer, S. Zeuzem, and J. Raedle. Mutations in the MutSalpha interaction interface of MLH1 can abolish DNA mismatch repair. *Nucleic Acids Research*, 34(22):6574–6586, 2006.

[43] J. Rahnenführer. Clustering algorithms and other exploratory methods for microarray data analysis. *Methods of Information in Medicine*, 44(3):444–448, 2005.

[44] J. Rahnenführer. Image analysis for cDNA microarrays. *Methods of Information in Medicine*, 44(3):405–407, 2005.

[45] 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.

[46] 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.

[47] M. Ralser, M. Albrecht, U. Nonhoff, T. Lengauer, H. Lehrach, and S. Krobitsch. An integrative approach to gain insights into the cellular function of human ataxin-2. *Journal of Molecular Biology*, 346(1):203–214, 2005.

[48] M. Ralser, U. Nonhoff, M. Albrecht, T. Lengauer, E. E. Wanker, H. Lehrach, and S. Krobitsch. Ataxin-2 and huntingtin interact with endophilin-A complexes to function in plastin-associated pathways. *Human Molecular Genetics*, 14(19):2893–2909, 2005.

[49] 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.

[50] 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. In U. Leser, F. Naumann, and B. Eckman, eds., *Data Integration in the Life Sciences, Third International Workshop, DILS 2006*, July 2006, *LNCS 4075*, pp. 185–194. Springer.

[51] N. Salamat-Miller, J. Fang, C. W. Seidel, Y. Assenov, M. Albrecht, and R. C. Middaugh. A network-based analysis of polyanion-binding proteins utilizing human protein arrays. *Journal of Biological Chemistry*, 282(14):10153–10163, 2007.

[52] N. Salamat-Miller, J. Fang, C. W. Seidel, A. M. Smalter, Y. Assenov, M. Albrecht, and C. R. Middaugh. A network-based analysis of polyanion-binding proteins utilizing yeast protein arrays. *Molecular and Cellular Proteomics*, 5(12):2263–2278, 2006.

[53] 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):e58, 2007.

[54] O. Sander, I. Sommer, and T. Lengauer. Local protein structure prediction using discriminative models. *BMC Bioinformatics*, 7(14):1–13, January 2006.

[55] C. Sarrazin, U. Mihm, E. Herrmann, C. Welsch, M. Albrecht, U. Sarrazin, S. Traver, T. Lengauer, and S. Zeuzem. Clinical significance of in vitro replication-enhancing mutations of the hepatitis C virus (hcv) replicon in patients with chronic HCV infection. *The Journal of Infectious Diseases*, 192(10):1710–1719, 2005.

[56] 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, June 2006.

[57] 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, March 2007.

[58] S. Schreiber, P. Rosenstiel, M. Albrecht, J. Hampe, and M. Krawczak. Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nature Reviews Genetics*, 6(5):376–388, 2005.

[59] 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*, 6:14, 2007.

[60] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. ROCR: Visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, 2005.

[61] 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.

[62] I. Sommer, S. Toppo, O. Sander, T. Lengauer, and S. Tosatto. Improving the quality of protein structure models by selecting from alignment alternatives. *BMC Bioinformatics*, 7:1–11, July 2006.

[63] A. Steffen, A. Kämper, and T. Lengauer. Flexible docking of ligands into synthetic receptors using a two-sided incremental construction algorithm. *Journal of Chemical Information and Modeling*, 46(4):1695–1703, July 2006.

[64] V. Svicher, T. Sing, M. M. Santoro, F. Forbici, F. Rodriguez-Barrios, A. Bertoli, N. Beerenwinkel, M. C. Bellocchi, F. Gago, A. d´Arminio Monforte, A. Antinori, T. Lengauer, F. Ceccherini-Silberstein, and C. Perno. Involvement of novel HIV-1 reverse transcriptase mutations in the regulation of resistance to nucleoside inhibitors. *Journal of Virology*, 80(14):7186–7198, 2006.

[65] F. L. Takken, M. Albrecht, and W. I. Tameling. Resistance proteins: molecular switches of plant defence. *Current Opinion in Plant Biology*, 9(4):383–390, 2006.

[66] W. I. Tameling, J. H. Vossen, M. Albrecht, T. Lengauer, J. A. Berden, M. A. Haring, B. J. Cornelissen, and F. L. W. Takken. Mutations in the NB-ARC domain of I-2 that impair ATP hydrolysis cause autoactivation. *Plant Physiology*, 140(4):1233–1245, 2006.

[67] 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, P. László, 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 of the United States of America*, 104(13):5495–5500, 2007.

[68] 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, March 2005.

[69] 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, April 2005.

[70] R. Valentonyte, J. Hampe, K. Huse, P. Rosenstiel, M. Albrecht, A. Stenzel, M. Nagy, K. I. Gaede, A. Franke, R. Haesler, A. Koch, T. Lengauer, D. Seegert, N. Reiling, S. Ehlers, E. Schwinger, M. Platzer, M. Krawczak, J. Müller-Quernheim, M. Schürmann, and S. Schreiber. Sarcoidosis is associated with a truncating splice site mutation in BTNL2. *Nature Genetics*, 37(4):357–364, 2005.

[71] M. M. Van Duist, M. Albrecht, M. Podswiadek, D. Giachino, T. Lengauer, L. Punzi, and M. De Marchi. A new CARD15 mutation in Blau syndrome. *European Journal of Human Genetics*, 13(6):742–747, 2005.

[72] V. Velagapudi, C. Wittmann, T. Lengauer, P. Talwar, and E. Heinzle. Metabolic high- content screening of saccharomyces cerevisiae reveals superior performance of single gene deletion strains. *Process Biochemistry*, 41:2170–2179, 2006.

[73] M. Voets, I. Antes, C. Scherer, K. Biemel, C. Barassin, 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.

[74] 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.

[75] M. W. Welker, W.-P. Hofmann, C. Welsch, M. von Wagner, E. Herrmann, T. Lengauer, S. Zeuzem, and C. Sarrazin. Correlation of nonstructural (NS) 4B amino acid variations with initial viral kinetics during interferon alfa-based therapy in HCV 1b-infected patients. *Journal of Viral Hepatitis*, 14, 2007.

[76] C. Welsch, M. Albrecht, J. Maydt, E. Herrmann, M. W. Welker, C. Sarrazin, A. Scheidig, T. Lengauer, and S. Zeuzem. Structural and functional comparison of the non-structural protein 4B in flaviviridae. *Journal of Molecular Graphics and Modelling*, 25, 2007.

[77] S. Wemmert, R. Ketter, J. Rahnenführer, N. Beerenwinkel, M. Strowitzki, W. Feiden, C. Hartmann, T. Lengauer, F. Stockhammer, K. D. Zang, E. Meese, W.-I. Steudel, A. von Deimling, and S. Urbschat. Patients with high grade gliomas harboring delections of chromosomes 9p and 10q benefit from temozolomide treatment. *Neoplasia*, 7(10):883–893, 2005.

[78] J. Yin, N. Beerenwinkel, J. Rahnenführer, and T. Lengauer. Model selection for mixtures of mutagenetic trees. *Statistical Applications in Genetics and Molecular Biology*, 5(1):Article 17, 2006.

[79] H. Zhu, F. S. Domingues, I. Sommer, and T. Lengauer. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7:1–15, June 2006.

Conference articles

[1] A. Altmann, M. Däumer, T. Sing, N. Beerenwinkel, H. Walter, R. Kaiser, and T. Lengauer. Validation of geno2pheno and THEO on a large independent clinical dataset. In *Proceedings of the 5th European HIV Drug Resistance Workshop*, Paris, France, 2007, HIV Medicine. European AIDS Clinical Society.

[2] I. Antes and T. Lengauer. Refinement of homology modeled protein structures for molecular docking. In *Proceedings of the 367. WE-Heraeus Seminar on Biomolecular Simulations*, Bad Honnef, April 2006, vol. 1, p. 1. Wilhelm und Else Heraeus-Stiftung.

[3] I. Antes and T. Lengauer. Prediction of bound MHC-peptide conformations. In *Proceedings of the 11th Annual International Conference on Research on Computational Molecular Biology*, Oakland, May 2007, vol. 11. RECOMB.

[4] R. Kaiser, I. Boussaad, C. Lehmann, M. Däumer, T. Sing, N. Schmeisser, C. Wyen, and G. Fätkenheuer. Stable coreceptor usage of HIV-1 in patients with ongoing treatment failure on HAART. In *Proceedings of the 3rd European HIV Drug Resistance Workshop*, Athens, Greece, 2005, HIV Medicine. European AIDS Clinical Society.

[5] A. Kämper, M. Kaspar, and T. Lengauer. Efficient conformational analysis of synthetic receptors with macrocyclic and fused ring systems. In *Synthetic Receptors 2005: Second World Congress on Synthetic Receptors*, Salzburg, Austria, 2005, p. O4. Elsevier. Abstract for an oral presentation at the conference.

[6] A. Schlicker, C. Huthmacher, F. Ramírez, T. Lengauer, and M. Albrecht. Functional evaluation of domain-domain interactions and human protein interaction networks. In D. Huson, O. Kohlbacher, A. Lupas, K. Nieselt, and A. Zell, eds., *German Conference on Bioinformatics (GCB 2006)*, Tübingen, Germany, September 2006, *Lecture Notes in Informatics*, vol. P-83, pp. 115–126. Gesellschaft für Informatik.

[7] T. Sing and N. Beerenwinkel. Mutagenetic tree Fisher kernel improves prediction of HIV drug resistance from viral genotype. In S. B, P. J., and H. T., eds., *Advances in Neural Information Processing Systems 19*, Vancouver, B.C., Canada, 2007, pp. 1–9. MIT.

[8] T. Sing, V. Svicher, N. Beerenwinkel, F. Ceccherini-Silberstein, M. Däumer, R. Kaiser, H. Walter, K. Korn, D. Hoffmann, M. Oette, J. K. Rockstroh, G. Fätkenheuer, C.-F. Perno, and T. Lengauer. Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling and SVM-based feature ranking. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, eds., *Knowledge Discovery in Databases: PKDD 2005.*, Porto, Portugal, 2005, pp. 285–296. Springer.

[9] A. Steffen, A. Kämper, and T. Lengauer. FlexR - a new tool for predicting the structure of synthetic host-guest complexes. In A. Turner, ed., *Synthetic Receptors 2005: Second World Congress on Synthetic Receptors*, Salzburg, Austria, 2005, p. O7. Elsevier. Abstract for an oral presentation at the conference.

[10] A. Steffen, A. Kämper, and T. Lengauer. Virtual screening for guest molecules of a biosensor. In A. P. F. Turner, ed., *Biosensors 2006: The ninth world congress on biosensors*, Toronto, Canada, May 2006, p. O66. Elsevier.

[11] V. Svicher, F. Ceccherini-Silberstein, T. Sing, M. Santoro, C. Gori, N. Beerenwinkel, F. Gago, R. D'Arrigo, M. Bellocchi, S. Giannella, A. Bertoli, A. D'Arminio Monforte, A. Antinori, and C. Perno. Additional mutations in HIV-1 reverse transcriptase are involved in the highly ordered regulation of NRTI resistance. In *Proceedings of the 3rd European HIV Drug Resistance Workshop*, Athens, Greece, 2005, HIV Medicine. European AIDS Clinical Society.

[12] V. Svicher, F. Ceccherini-Silberstein, T. Sing, M. Santoro, C. Gori, N. Beerenwinkel, F. Gago, R. D'Arrigo, M. Bellochi, S. Giannella, A. Bertoli, A. D'Arminio Monforte, A. Antinori, and C. Perno. A highly ordered network of HIV-1 RT mutations regulates the continuous escape from antiviral drugs. In *International Workshop in HIV Dynamics and Evolution*, Cleveland, Ohio, USA, 2005. University of California, San Diego.

[13] P. Talwar, T. Lengauer, C. Wittmann, V. Velagapudi, and E. Heinzle. Development of computational methods for analysis of metabolic profiling data. In *Proceedings of International Conference on Systems Biology (ICSB 2006), Japan*, Yokohama, Japan, 2006, pp. 42–42. International Conference on Systems Biology 2006.

[14] M. Zazzi, E. Aharoni, A. Altmann, F. Bazsó, P. Bidgood, G. Borgulya, J. Denholm-Prince, M. Fielder, R. Kaiser, C. Kent, T. Lengauer, T. Nepusz, H. Neuvirth, Y. Peres, A. Petroczi, M. Prosperi, L. Romano, M. Rosen-Zvi, E. Schülter, T. Sing, A. Sonnerborg, R. Thompson, G. Ulivi, L. Zalány, and F. Incardona. Euresist: exploration of multiple modeling techniques for prediction of response to treatment. In *Proceedings of the 5th European HIV Drug Resistance Workshop*, Paris, France, 2007, HIV Medicine. European AIDS Clinical Society.

Theses

[1] M. Albrecht. *Combining protein structure prediction with experiments and functional information.* Phd thesis, Universität des Saarlandes, February 2006.

[2] A. Alexa. Integrating the GO graph structure in scoring the significance of gene ontology terms. Masters thesis, Universität des Saarlandes, March 2005.

[3] T. Binsl. Feature prints: A new method for calculating molecular similarities. Masters thesis, Universität des Saarlandes, 2006.

[4] J. Bogojeska. Stability analysis of oncogenetic trees mixture models. Masters thesis, Universität des Saarlandes, January 2007.

[5] Y. Djoumbou. Docking studies to investigate the effect of correlated mutations in the HIV-protease on the drug resistance of HIV. Bachelor thesis, Universität des Saarlandes, December 2005.

[6] O. Frings. Development of a smith-waterman-like multi-aspect alignment tool. Bachelor thesis, Universität des Saarlandes, January 2006.

[7] E. Glaab. On the predictability of cpg methylation in human tissue at single basepair resolution. Bachelor thesis, Universität des Saarlandes, February 2006.

[8] X. Guo. Improving fold recognition using protein and domain interactions. Masters thesis, Universität des Saarlandes, June 2006.

[9] K. Halachev. Epigraphregression: A toolkit for (epi-)genomic correlation analysis and prediction of quantitative attributes. Masters thesis, Universität des Saarlandes, September 2006.

[10] A. Hobler. Docking of monosaccharids into the wild type and mutants of pyranose 2-oxidase. Bachelor thesis, Universität des Saarlandes, June 2006.

[11] C. Huthmacher. Decomposing protein networks into domain-domain interactions. Diploma thesis, Universität des Saarlandes, May 2006.

[12] S. Immesberger. A prefiltering method to speed up protein structure prediction. Bachelor thesis, Universität des Saarlandes, January 2005.

[13] H. Jung. Yamada analyzer - a tool for fast calculation of genomic attributes. Bachelor thesis, Universität des Saarlandes, December 2006.

[14] M. Kaspar. Conformational analysis of macrocyclic ring systems. Bachelor thesis, Universität des Saarlandes, January 2005.

[15] M. Kaspar. Automated conformational analysis of macrocyclic ring systems. Masters thesis, Universität des Saarlandes, July 2006.

[16] M. Krier. Motif search in protein sequences for protein structure prediction. Bachelor thesis, Universität des Saarlandes, June 2005.

[17] I. Meiser. Identification of functionally important regions in cytochrome P450 systems and a systematic analysis of various CYP P450 homology models. Bachelor thesis, Universität des Saarlandes, July 2005.

[18] O. Mueller. Using shape retrieval techniques for identifying similar protein binding sites. Masters thesis, Universität des Saarlandes, September 2007.

[19] S. Nickels. Generating molecular hashcodes for finding duplicates in virtual screening libraries. Bachelor thesis, Universität des Saarlandes, September 2005.

[20] S. Pfeifer. Development of a fast geometry-based screening compound prefiltering technique. Bachelor thesis, Universität des Saarlandes, September 2006.

[21] J. Rahnenführer. *Statistical methods for the biological interpretation of genome-wide measurements*. Habilitation thesis, Universität des Saarlandes, October 2006.

[22] D. Reimer. Docking studies to investigate the effect of correlated mutations in the HIV-Protease on the drug resistance of HIV. Bachelor thesis, Universität des Saarlandes, June 2006.

[23] A. Schlicker. A global approach to comparative genomics: Comparison of functional annotation over the taxonomic tree. Masters thesis, Universität des Saarlandes, August 2005.

[24] W.-I. Siu. Computational prediction of MHC-peptide interactions. Masters thesis, Universität des Saarlandes, September 2005.

[25] A. Thielen. Prediction of linear B-cell epitopes and MHC class II binding peptides. Masters thesis, Eberhardt-Karls-Universität Tübingen, March 2006.

[26] L. Tolosi. Analysis of array CGH data for the estimation of genetic tumor progression. Masters thesis, Universität des Saarlandes, June 2006.

[27] A. Volkamer. Implementation of a method to filter out compounds with toxic, reactive and unsuitable functional groups. Bachelor thesis, Universität des Saarlandes, April 2005.

[28] N. Weinhold. Inference of protein function based on functionally conserved regions in sequence and structure space. Masters thesis, Universität des Saarlandes, December 2006.

[29] J. Weiss. Generating frequency profiles from multiple structure alignments. Bachelor thesis, Universität des Saarlandes, June 2005.

[30] G. Yalak. Analysis of mutations in the death domains, death effector domains, caspase recruitment domains, and pyrin domains. Bachelor thesis, Universität des Saarlandes, February 2005.

[31] J. Yin. Model selection for mixtures of mutagenetic trees. Masters thesis, Universität des Saarlandes, March 2005.

[32] T. Zimmer. Functional site scaffolds. Bachelor thesis, Universität des Saarlandes, June 2006.

15 The Computer Graphics Group (D4)

15.1 Personnel

Director

Prof. Dr. Hans-Peter Seidel

Researchers

Dr. Alexander Belyaev
Dr. Volker Blanz (–September 2005)
Dr. Kirill Dmitriev (–Mai 2005)
Dr. Michael Goesele (–August 2005)
Dr. Stefan Gumhold (–October 2005)
Dr. Jörg Haber (–September 2005)
Dr. Vlastimil Havran (–January 2006)
Dr. Ioannis Ivrissimtzis (–September 2005)
Dr. Zachi Karni (–September 2006)
Dr. Hendrik Lensch (April 2006–)
Dr. Rafal Mantiuk
Dr. Karol Myszkowski
Dr. Bodo Rosenhahn (–December 2005)
Dr. Christian Rössl (–December 2005)
Dr. Holger Theisel (–November 2006)
Dr. Christian Theobalt (–April 2007)
Dr. Hitoshi Yamauchi (–June 2006)

PhD Students

Naveed Ahmed
Boris Ajdin (September 2006–)
Irene Albrecht (–August 2006)
Thomas Annen
Tunc Aydin (April 2006–)
Robert Bargmann
Tongbo Chen
Edilson De Aguiar
Zhao Dong
Alexander Efremov (–December 2005)
Christian Fuchs
Martin Fuchs

Jürgen Gall (January 2006–)
Mardé Greeff (–September 2005)
Johannes Günther
Nils Hasler (July 2006–)
Robert Herzog (November 2005–)
Matthias Hullin (April 2007–)
Grzegorz Krawczyk
Torsten Langer
Waqar Saleem
Natascha Sauber (September 2006–)
Oliver Schall
Kristina Scherbaum (September 2006–)
Thomas Schultz (May 2006–)
Kuangyu Shi
Kaleigh Smith (June 2005–)
Wenhao Song (July 2005–)
Carsten Stoll
Martin Sunkel (Februrary 2006–)
Wolfram von Funck
Akiko Yoshida
Shin Yoshizawa (–December 2006)
Rhaleb Zayer
Gernot Ziegler

Project Coordination
Dr. Christel Weins

Secretaries
Sabine Budde
Cornelia Liegl

15.2 Visitors

Since May 2005, a total of 83 researchers have visited our group.

Martin Isenburg	01.02.05–30.06.05	University of North Carolina, Chapel Hill, USA
Isabelle Sivignon	21.03.05–31.07.05	Institut National Polytechnique de Grenoble, Grenoble, France
Marc Spoor	26.04.05	Universität Kaiserslautern, Kaiserslautern, Germany
Guiseppe Patanè	30.04.05–29.06.05	IMATI, Genoa, Italy
Wenping Wang	17.05.05–28.05.05	The University of Hong Kong, Hong Kong, China

Victoria Hernandez	23.05.05–28.05.05	ICIMAF, La Habana, Cuba
Bernd Hoefflinger	06.06.05	IMS Chips, Stuttgart, Germany
David Brosch	06.06.05	IMS Chips, Stuttgart, Germany
Verena Schneider	06.06.05	IMS Chips, Stuttgart, Germany
Volker Gengenbach	06.06.05	IMS Chips, Stuttgart, Germany
Hans Kober	06.06.05	IMS Chips, Stuttgart, Germany
Dr. Roland Höfling	28.07.05	ViALUX, Chemnitz, Germany
Yaron Lipman	11.08.05–18.08.05	Tel-Aviv University, Tel-Aviv, Israel
Sung Yong Shin	01.09.05–31.10.05	KAIST, Guseong-dong Yuseong-gu Deajon, South Korea
Bodo Rosenhahn	08.09.05–14.09.05	University of Auckland (CITR), Auckland, New Zealand
Scott Daly	16.09.05–17.09.05	Sharp Laboratories of America, Camas, USA
Min Chen	19.09.05	University of Wales Swansea, Swansea, UK
Michael Wand	05.11.05–19.11.05	Stanford University, Stanford, USA
Leonidas Guibas	10.11.05–19.11.05	Stanford University, Stanford, USA
Billy Chen	11.11.05–19.11.05	Stanford University, Stanford, USA
Gaurav Garg	11.11.05–19.11.05	Stanford University, Stanford, USA
Hendrik Lensch	11.11.05–19.11.05	Stanford University, Stanford, USA
Aditya Mavlankar	11.11.05–16.11.05	Stanford University, Stanford, USA
David Varodayan	11.11.05–16.11.05	Stanford University, Stanford, USA
Joyce Farrell	12.11.05–20.11.05	Stanford University, Stanford, USA
Natasha Gelfand	12.11.05–16.11.05	Stanford University, Stanford, USA
Nikola Milosavljevic	12.11.05–08.01.06	Stanford University, Stanford, USA
Edward Adelson	13.11.05–15.11.05	Massachusetts Institute of Technology, Cambridge, USA
Volker Blanz	13.11.05–16.11.05	Universität Siegen, Siegen, Germany

Chuo-Ling Chang	13.11.05–16.11.05	Stanford University, Stanford, USA
Markus Flierl	13.11.05–16.11.05	Stanford University, Stanford, USA
Bernd Girod	13.11.05–16.11.05	Stanford University, Stanford, USA
Andrew Selle	13.11.05–16.11.05	Stanford University, Stanford, USA
Robert Strzodka	13.11.05–16.11.05	Stanford University, Stanford, USA
Thomas Ertl	14.11.05–15.11.05	Universität Stuttgart, Stuttgart, Germany
Eugene Fiume	14.11.05–16.11.05	University of Toronto, Toronto, Kanada
Stefan Gumhold	14.11.05–16.11.05	Technische Universität Dresden, Dresden, Germany
Jörn Ostermann	14.11.05–16.11.05	Universität Hannover, Hannover, Germany
Roberto Scopigno	14.11.05–16.11.05	Consiglio Nazionale delle Ricerche (CNR), Pisa, Italy
Martin Vetterli	14.11.05–15.11.05	École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
Volker Blanz	22.12.05–23.12.05	Universität Siegen, Siegen, Germany
Kai Harmann	06.01.06–07.01.06	Technische Universität Clausthal, Clausthal-Zellerfeld, Germany
Daniel Grest	04.02.06–10.02.06	Christian-Albrechts-Universität Kiel, Kiel, Germany
Carsten Dachsbacher	14.02.06–15.02.06	Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
Dorota Zdrojewska	01.03.06–13.04.06	Szczecin University of Technology, Szczecin, Poland
Gabriel Fournier	13.03.06–15.03.06	Université Claude Bernard, Lyon, France
Jack Tumblin	19.03.06–21.03.06	Northwestern University, Evanston, USA
Xavier Pueyo	29.03.06–30.03.06	Universitat de Girona, Girona, Italy
Peter Longhurst	03.04.06–10.04.06	University of Bristol, Bristol, UK

Boris Ajdin	08.04.06–09.04.06	University of Belgrade, Belgrade, Serbia-Montenegro
Gabriel Fournier	10.04.06–31.08.06	Université Claude Bernard, Lyon, France
Wolfgang Heidrich	10.04.06–13.04.06	University of British Columbia, Vancouver, Canada
Markus Gross	02.05.06	ETH Zürich, Zürich, Switzerland
Laurent Kirsche	10.05.06	PSA Peugeot Citroen, Paris, France
Philippe Porral	10.05.06	PSA Peugeot Citroen, Paris, France
Pierpaolo Baccichet	16.05.06	Universitá degli studi di Milano, Milano, Italy
Mario Capitelli	17.06.06	European Aeronautic Defence and Space Company (EADS), Hamburg
Juedith Daerr	17.06.06	European Aeronautic Defence and Space Company (EADS, Hamburg)
Barosz Spitzbarth	17.06.06	European Aeronautic Defence and Space Company (EADS), Hamburg
Wojciech Jaskowski	18.05.06–19.05.06	Poznan University of Technology, Poznan, Poland
Tsunoyamna	28.05.06–30.05.06	University of Aizu, Aizu, Japan
Reinhard Klette	03.07.06–04.07.06	University of Auckland, Auckland, New Zealand
Horst Zuse	14.07.06	FH Lausitz, PD TU Berlin
Meinard Müller	16.08.06–18.08.08	Bonn University, Bonn, Germany
Andrew Nealen	16.08.06–17.08.06	Technische Universität Berlin, Berlin, Germany
Olga Sorkine	16.08.06–17.08.06	Technische Universität Berlin, Berlin, Germany
Marc Pollefeys	21.08.06–22.08.06	University of North Carolina, Chapel Hill, USA
Jan-Michael Frahm	17.09.06–18.09.06	University of North Carolina, Chapel Hill, USA
Vladimir Garanzha	28.09.06–01.10.06	Computing Center Russian Academy of Sciences, Moscow, Russia

Mario Botsch	11.10.06–13.10.06	ETH Zurich, Zurich, Switzerland
Volker Blanz	12.10.06–14.10.06	Universität Siegen, Siegen, Germany
Hartmut Schirrmacher	12.10.06–14.10.06	Mercury Computer Systems GmbH, Berlin, Germany
Yutaka Ohtake	02.11.06–23.11.06	University of Aizu, Aizu, Japan
Jens Krüger	06.11.06–07.11.06	Technische Universität München, München, Germany
Sebastian Haering	07.11.06	Volkswagen AG, Wolfsburg, Germany
Eduard Jundt	07.11.06	Volkswagen AG, Wolfsburg, Germany
Nikolay Mirenkov	16.10.06–17.10.06	University of Aizu, Aizu, Japan
Xavier Pueyo	29.11.06	Universitat de Girona, Girona, Italy
Stephan Würmlin	29.11.06–30.11.06	ETH Zürich, Zurich, Switzerland
Michael Waschbüsch	30.11.06–01.12.06	ETH Zürich, Zurich, Switzerland
Xinjie Yu	28.11.06–30.11.06	Tsinghua University, Tsinghua, China
Radoslaw Mantiuk	11.02.07–06.04.07	Szczecin Unviersity, Szczecin, Poland
Simon Winkelbach	26.02.07–27.02.07	Universität Braunschweig, Braunschweig, Germany
Sergey Tsarev	05.03.07–07.07.07	TU Berlin, Berlin, Germany

15.3 Group Organization

Our research is currently organized into the following nine research areas, each having its own small group of coordinators:

- Digital Geometry Processing (A. Belyaev)
- Freeform Surfaces and Visualization (C. Rössl)
- Vector Field Visualization and Applications of Vector Field Techniques (H. Theisel)
- Modeling and Animation of Faces (V. Blanz)
- Markerless Motion Capture (B. Rosenhahn)
- Multiview Video Processing (C. Theobalt)
- General Appearance Acquisition and Computational Photography (H. Lensch)

- Advanced Global Illumination and Realtime Realistic Image Synthesis (J. Günther and K. Myszkowski)

- High Dynamic Range Imaging and Perception Issues in Graphics (R. Mantiuk and K. Myszkowski)

The coordinators coordinate the work in their areas and together with Hans-Peter Seidel 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, 1pm), 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, 12:00 pm), 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.

15.4 Digital Geometry Processing

Coordinator: Alexander Belyaev

We are witnessing an explosion in the use of digital media: digital sound, image, video, and since very recently – digital geometry. Rapid advances in 3D shape acquisition technologies are forcing fast and impressive development of Digital Geometry Processing, a new research area whose goal is to bring new mathematical and computational tools needed for efficient processing of 3D geometry information. Being typical features of 3D geometric data, curved geometry, nontrivial topology, irregular sampling, and huge data size prevent a straightforward adaptation of signal processing techniques and pose new mathematical and algorithmic challenges.

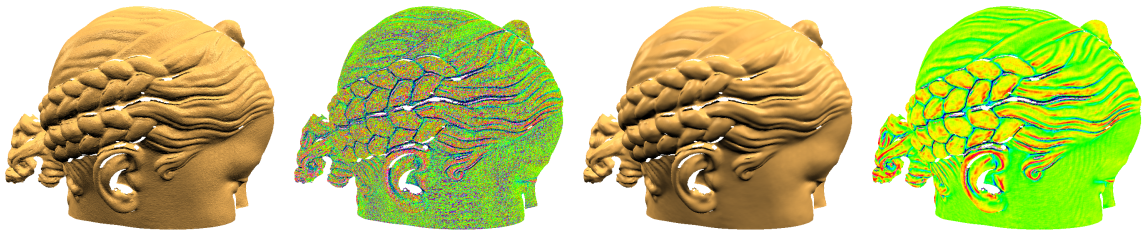


Figure 15.1: The left pair of images shows a laser range scan of the Bimba model and its corresponding mean curvature visualization. It is clearly visible that the scan is corrupted by dense high-frequency noise. The right pair of images illustrates the denoising result of our algorithm. Noise is properly removed while important details are preserved.

15.4.1 Shape Denoising, Reconstruction, and Representation

Investigators: Oliver Schall, Carsten Stoll, Waqar Saleem, Ioannis Ivrissimtzis, Christian Rössl, Hitoshi Yamauchi, Stefan Gumhold, Alexander Belyaev, Zachi Karni, and Hans-Peter Seidel

Modern 3D digitizing techniques can yield millions of 3D point locations on the object that is being digitized. Once these points have been collected, it is a non-trivial task to build a surface representation that is faithful to the collected data. Some of the desirable properties of a surface reconstruction method include high speed, low memory overhead, faithful reproduction of delicate surface features, and robustness in the presence of noise, holes, and variations in sampling density. Our research in this field focuses on developing robust and accurate analytical, stochastic, and learning approaches for high fidelity reconstruction and representation of 3D objects from scattered data.

Feature-preserving surface denoising. Real-world signals do not exist without noise. Denoising digital images and their 3D geometry counterparts, polygonal meshes and point clouds, remains an active area of research. Our research in this filed is focused on developing non-local filtering schemes for feature-preserving denoising of static [11, 19] (see Figure 15.1 for an example of feature-preserving denoising of a model with many fine details) and time-varying [13, 15] geometric data. We also show that while some established meshing techniques often fail to reconstruct the surface from original noisy point scattered data, they work well in conjunction with the kernel based clustering approach introduced in [11].

Learning techniques for surface reconstruction. In [3], we study the use of ensembles in surface reconstruction. By randomly subsampling the input data we create several surface reconstructions, which are then combined into a single surface with better qualities than any individual member of the ensemble. In [4, 18] we further study the use of ensembles in surface reconstruction and normal estimation. In particular, we measure the effect of different sampling and averaging methods on the quality of the ensemble, especially in the presence of outlier noise. Due to the complexity of the underlying surface reconstruction and normal estimation algorithms, our results are based on extensive experimentation with artificial and

scan data. In [2], we propose an algorithm for overfitting control in surface reconstruction, based on an octree partition of the training samples. At each level of the octree, we use one representative sample from each leaf cell and reconstruct a surface. Then, we test each leaf cell for overfitting and split only those cells where overfitting has not been detected.

Adaptive surface fitting. Adaptively refined composite implicit surfaces provide us with extremely useful tools for interpolation and approximation of scattered 3D data. In [12, 14] a spatial error-guided subdivision is combined with a FFT-based surface reconstruction technique. Adaptive surface fitting with radial basis functions and polynomial local approximations is a subject of research in [6, 9, 5]. Several combinations of implicit surface fitting techniques and statistical methods for surface reconstruction are studied in [10]. In [17] a simple implicit representation of a surface is achieved using a BSP-tree construction: a continuous surface approximation is represented as a single unified BSP-tree structure, with the leaf cells representing a local surface patch, allowing for easy storage and compression.

Template-based surface reconstruction. Conventional surface reconstruction techniques often fail to deliver a correct reconstruction of a 3D object from severely incomplete scattered data with a large number of outliers. In order to address this limitation, we propose in [16] a method which uses an initial template model of a similar object to the one being reconstructed as prior knowledge. The main idea of our approach is to deform the template model guided by a low number of user specified correspondence points to fit to the input points. This is achieved by first deforming the template to a rough pose of the object using a Laplacian deformation and afterwards iteratively adjusting the shape to fit to the input data as close as possible using surface subdivision to maintain good sample density, a projection operator generating new constrained positions for the template close to the input data and a Laplacian deformation ensuring a high quality surface.

A spherical cover approach to meshing scattered surface data. In [7, 8] we propose a new method for approximating an unorganized set of non-oriented points scattered over a piecewise smooth surface by a triangle mesh. The method consists of three stages. First an adaptive spherical cover and auxiliary points corresponding to the cover elements are generated. Then the intersections between the spheres of the cover are analyzed and the auxiliary points are connected. Finally the resulting mesh is cleaned from non-manifold parts. The method allows us to control the approximation accuracy, is capable of processing noisy data, and shows a good performance is reconstructing sharp edges and corners.

Error-controlled LOD representation of polygonal models. A level-of-detail representation (LOD) is a key geometry processing technique for an efficient handling of huge polygonal models. In [1] a new mesh hierarchy was introduced that builds on edge-removal/insert and edge-join/split operations. It allows for the first time truly selective refinement with control of the two-sided Hausdorff distance.

References

- [1] S. Gumhold. Truly selective polygonal mesh hierarchies with error control. *Computer Aided Geometric Design*, 22(5):424–443, 2005.
- [2] Y. Lee, S. Lee, I. Ivrissimtzis, and H.-P. Seidel. Overfitting control for surface reconstruction. In D. W. Fellner, S. N. Spencer, A. Sheffer, and K. Polthier, eds., *SGP 2006: Fourth Eurographics Symposium on Geometry Processing*, Cagliari, Sardinia, Italy, June 2006, pp. 231–234. Eurographics.
- [3] Y. Lee, M. Yoon, S. Lee, I. Ivrissimtzis, and H.-P. Seidel. Ensembles for surface reconstruction. In C. Gotsman, D. Manocha, and E. Wu, eds., *Proceedings of the 13th Pacific Conference on Computer Graphics and Applications*, Macao, October 2005, pp. 125–127. Welfare Printing Ltd.
- [4] Y. Lee, M. Yoon, S. Lee, I. Ivrissimtzis, and H.-P. Seidel. Ensembles for normal and surface reconstructions. In *Geometric Modeling and Processing (GMP 2006)*, Pittsburgh, Pennsylvania, USA, 2006, *LNCS 4077*, pp. 17–33. Springer.
- [5] Y. Ohtake, A. Belyaev, and M. Alexa. Sparse low-degree implicit surfaces with applications to high quality rendering, feature extraction, and smoothing. In M. Desbrun and H. Pottman, eds., *Eurographics Symposium on Geometry Processing 2005*, Vienna, Austria, 2005, pp. 149–158. Eurographics.
- [6] Y. Ohtake, A. Belyaev, and H.-P. Seidel. 3d scattered data interpolation and approximation with multilevel compactly supported rbfs. *Graphical Models*, 67(3):150–165, 2005.
- [7] Y. Ohtake, A. Belyaev, and H.-P. Seidel. An integrating approach to meshing scattered point data. In *ACM Symposium on Solid and Physical Modeling (SPM 2005)*, Cambridge, MA, USA, June 2005, pp. 61–69. ACM.
- [8] Y. Ohtake, A. Belyaev, and H.-P. Seidel. A composite approach to meshing scattered data. *Graphical Models*, 68(3):255–267, 2006.
- [9] Y. Ohtake, A. Belyaev, and H.-P. Seidel. Sparse surface reconstruction with adaptive partition of unity and radial basis functions. *Graphical Models*, 68(1):15–24, January 2006.
- [10] W. Saleem, O. Schall, G. Patanè, A. Belyaev, and H.-P. Seidel. On stochastic methods for surface reconstruction. *The Visual Computer*, XX, 2007. to appear.
- [11] O. Schall, A. Belyaev, and H.-P. Seidel. Robust filtering of noisy scattered point data. In M. Pauly and M. Zwicker, eds., *IEEE/Eurographics Symposium on Point-Based Graphics*, Stony Brook, New York, USA, 2005, pp. 71–77. Eurographics.
- [12] O. Schall, A. Belyaev, and H.-P. Seidel. Adaptive fourier-based surface reconstruction. In M.-S. Kim and K. Shimada, eds., *Geometric modeling and processing - GMP 2006, 4th International Conference*, Pittsburgh, PA, USA, 2006, *LNCS 4077*, pp. 34–44. Springer.
- [13] O. Schall, A. Belyaev, and H.-P. Seidel. Feature-preserving denoising of time-varying range data. In H. Pfister, ed., *SIGGRAPH 2006 Sketches and Applications*, Boston, Massachusetts, USA, 2006, p. 56. ACM.
- [14] O. Schall, A. Belyaev, and H.-P. Seidel. Error-guided adaptive fourier-based surface reconstruction. *Computer-Aided Design*, XX, 2007. to appear.
- [15] O. Schall, A. Belyaev, and H.-P. Seidel. Feature-preserving non-local denoising of static and time-varying range data. In *ACM Symposium on Solid and Physical Modeling*, Beijing, China, 2007. ACM. to appear.

- [16] C. Stoll, Z. Karni, C. Rössl, H. Yamauchi, and H.-P. Seidel. Template deformation for point cloud fitting. In M. Botsch and B. Chen, eds., *Symposium on Point-Based Graphics*, Boston, USA, 2006, pp. 27–35. Eurographics.
- [17] C. Stoll, H.-P. Seidel, and M. Alexa. Bsp shapes. In *2006 International Conference on Shape Modeling and Applications (SMI 2006)*, Matsushima, Japan, June 2006, Proceedings of the IEEE International Conference on Shape Modeling and Applications, pp. 42–47. IEEE.
- [18] M. Yoon, Y. Lee, S. Lee, I. Ivrissimtzis, and H.-P. Seidel. Surface and normal ensembles for surface reconstruction. *Computer-Aided Design*, XX, 2007.
- [19] S. Yoshizawa, A. Belyaev, and H.-P. Seidel. Smoothing by example: Mesh denoising by averaging with similarity-based weights. In M. Spagnuolo, A. Belyaev, and H. Suzuki, eds., *IEEE International Conference on Shape Modeling and Applications 2006 (SMI 2006)*, Matsushima, JAPAN, June 2006, pp. 38–44. IEEE.

15.4.2 Discrete Differential Geometry

Investigators: Rhaleb Zayer, Carsten Stoll, Torsten Langer, Shin Yoshizawa, Stefan Gumhold, Zachi Karni, Christian Rössl, Hitoshi Yamauchi, and Alexander Belyaev

Discrete differential geometry is a new research area emerging at the intersection of discrete and differential geometry and approximation theory. It aims at developing computationally efficient discrete counterparts of concepts and methods of continuous differential geometry. Our research in this area is focused on developing methods for low-distortion discrete surface parameterization, convergence analysis of discrete differential operators, and data interpolation with generalized barycentric coordinates.

Discrete surface parameterization. Surface parameterization consists of a surface decomposition into a set of patches, also referred to as an atlas of charts, and establishing one-to-one mappings between the patches and reference domains. Numerous applications of surface parameterization in computer graphics and geometric modeling include texture mapping, shape morphing, surface reconstruction and repairing, and grid generation.

In [5] we propose a new method for a decomposition of a given surface into a set of charts that can be flattened efficiently. Single-chart surface parameterization approaches are studied in [6, 7, 8]. In particular, a simple and efficient boundary-free mesh parameterization approach is proposed in [7]. A single-chart spherical parameterization is studied in [8] where the use of appropriately chosen coordinates allows for an approximation a nonlinear mesh parameterization problem by a linear one and delivers a fair balance between high-quality and computational efficiency(see Figure 15.2).

While traditional mesh parameterization algorithms tries to find a planar embedding of a given surface, in [4] we also investigate the reverse method in the context of constrained texture parameterization, deforming the texture plane to fit the surface in 3D using a Laplacian deformation based on geodesic distances from the constraints.

Curvature tensor estimation. An accurate and robust estimation of curvature properties of a smooth surface approximated by a polygonal mesh is important for many geometric modeling and computer graphics applications. In [3] we present a novel approach to calculate

Figure 15.2: The left pair of images shows a free boundary planar parameterization of a gargoyle model. The right pair of images shows the spherical parameterization of a turtle model. Both results illustrate the robustness and low distortion qualities of the proposed methods.

the mean curvature from arbitrary normal curvatures. Then, we demonstrate how the same method can be used to obtain new formulae to compute the Gaussian curvature and the curvature tensor. The idea is to compute the curvature integrals by a weighted sum by making use of the periodic structure of the normal curvatures to make the quadratures exact. Finally, we derive an approximation formula for the curvature of discrete data like meshes and show its convergence if quadratically converging normals are available.

Generalized barycentric coordinates. Generalized barycentric coordinates are widely used in computational mechanics and fluid dynamics where numerical methods based on barycentric interpolation schemes (the so-called meshless finite element methods and finite volume methods) gain increasing popularity. Boundary data interpolation and free-form shape deformations constitute other application areas for barycentric coordinates and their generalizations. In [2] we develop spherical barycentric coordinates and use them to extend the so-called mean value coordinates to arbitrary polyhedra. In our approach, we reduced the problem of finding spherical barycentric coordinates to the planar case by using the gnomonic (or central) projection from the sphere to a specific tangent plane. We observe that this projection establishes a bijective correspondence between planar and spherical barycentric coordinates that preserves many properties of the respective coordinates. One of the most interesting consequences is the possibility to construct 3D mean value coordinates for arbitrary polyhedra. We show that the spherical and 3D mean value coordinates are well defined on the whole sphere and the whole 3D space, respectively. Figure 15.3 demonstrates some immediate advantages of our generalization of the 3D mean value coordinates.

In [1] we present a general construction of transfinite barycentric coordinates and reveal their relations with classical inverse problems in differential and convex geometry.

References

- [1] A. Belyaev. On transfinite barycentric coordinates. In A. Sheffer and K. Polthier, eds., *SGP 2006, Fourth Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, Cagliari, Sardinia,

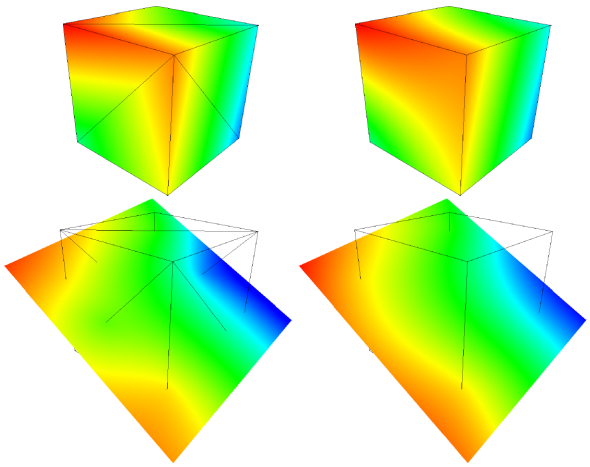


Figure 15.3: An example of interpolation of color values using 3D mean value coordinates. The color values are specified at the vertices of the cube. They are interpolated on the faces (top) and on a plane passing through the cube (bottom). If the cube is triangulated beforehand (left), the interpolation is less smooth than with our method (right).

Italy, 2006, pp. 89–99. Eurographics.

- [2] T. Langer, A. Belyaev, and H.-P. Seidel. Spherical barycentric coordinates. In D. W. Fellner, S. N. Spencer, A. Sheffer, and K. Polthier, eds., *SGP 2006, Fourth Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, Cagliari, Sardinia, Italy, June 2006, pp. 81–88. Eurographics.

- [3] T. Langer, A. Belyaev, and H.-P. Seidel. Exact and interpolatory quadratures for curvature tensor estimation. *Computer Aided Geometric Design*, Accepted, 2007.

- [4] C. Stoll, Z. Karni, and H.-P. Seidel. Geodesics guided constrained texture deformation. In *The 14th Pacific Conference on Computer Graphics and Applications Proceedings*, Taipei, Taiwan, October 2006, *Pacific Conference on Computer Graphics and Applications Proceedings*, vol. 14, pp. 144–152. National Taiwan University Press.

- [5] H. Yamauchi, S. Gumhold, R. Zayer, and H.-P. Seidel. Mesh segmentation driven by gaussian curvature. *The Visual Computer*, 21(8-10):649–658, September 2005.

- [6] S. Yoshizawa, A. Belyaev, and H.-P. Seidel. A moving mesh approach to stretch-minimizing mesh parameterization. *International Journal of Shape Modeling*, 11(1):25–42, June 2005.

- [7] R. Zayer, C. Rössl, and H.-P. Seidel. Setting the boundary free: A composite approach to surface parameterization. In M. Desbrun and H. Pottmann, eds., *Symposium on Geometry Processing*, Vienna, Austria, 2005, pp. 91–100. Eurographics/ACM.

- [8] R. Zayer, C. Rössl, and H.-P. Seidel. Curvilinear spherical parameterization. In *Shape Modeling International (SMI)*, Matsushima, Japan, 2006, pp. 57–64. IEEE.

15.4.3 Shape Processing for Storage and Transmission

Investigators: Hitoshi Yamauchi, Waqar Saleem, Shin Yoshziawa, Stefan Gumhold, Zachi Karni, Martin Isenburg, and Alexander Belyaev

With the rapid growth of Internet technology, digital libraries become a major source of information and data for scientists and general public. In particular, the amount of 3D geometric information available online is growing dramatically. This poses new problems in compressing and automatic annotating 3D geometric data for transmission and storage applications.

Compression of polygonal meshes with high degree polygons. Limited transmission bandwidth hampers working with large polygon meshes in networked environments. This has motivated researchers to develop compressed representations for such data sets. While traditionally mesh compression schemes have focused on purely triangular meshes, many meshes found in 3D model libraries or output from modeling packages contain often only a small percentage of triangles. The dominating element is generally the quadrilateral, but pentagons, hexagons and higher degree faces are also common. In [3] we present a general method for extending the popular parallelogram geometry prediction rule for more efficient predictions within pentagons, hexagons, and other high degree polygons. Our method is based on the decomposition of the polygons into the Fourier basis and the prediction that missing points will be in a position that does not introduce high frequencies.

Point-cloud compression. Point clouds have recently become a popular alternative to polygonal meshes for representing three-dimensional geometric models. With modern scanning technologies producing larger and larger amounts of point data, it is beneficial to have compact representations for storage and transmission of such data. In [1, 2] we present a novel predictive method for single-rate compression of 3D models represented as point-clouds. Our method constructs a prediction tree that specifies which previously encoded points are used for predicting the position of the next point. We use a greedy point-by-point construction of the tree that tries to minimize the prediction error. The results show that our approach can compete with other schemes for compressing point positions. Because our method can be adapted for streaming, out-of-core operation, we can compress arbitrary large point sets using only moderate memory resources.

Shape annotating and view selection. The nature of 3D shape perception has intrigued humans for many centuries, and continues even today to be a subject of intensive research. In [4] we study the following problem which recently attracted a considerable attention of the pattern recognition and graphics communities. Given a 3D model, find view directions which deliver representative views of that object. Our interest in representative views is mainly in context of shape repositories, where information on the shape of stored models has to be presented to users in the form of a few representative views. In [4] we present a new view selection approach based on two fundamental elements of human perception: view stability and saliency.

References

- [1] S. Gumhold, Z. Karni, M. Isenburg, and H.-P. Seidel. Predictive point-cloud compression. In *SIGGRAPH 2005 Technical Sketches*, Los Angeles, USA, 2005. ACM.

- [2] S. Gumhold, Z. Karni, M. Isenburg, and H.-P. Seidel. Predictive point-cloud compression. In *Proceedings of the Sixth Israel-Korea Bi-National Conference*, Haifa, Israel, 2005, pp. 125–129. Technion.

- [3] M. Isenburg, I. Ivrissimtzis, S. Gumhold, and H.-P. Seidel. Geometry prediction for high degree polygons. In B. Jüttler, ed., *21st Spring Conference on Computer Graphics (SCCG 2005)*, Budmerice, Slovakia, 2005, pp. 147–152. Comenius University.

- [4] H. Yamauchi, W. Saleem, S. Yoshizawa, Z. Karni, A. Belyaev, and H.-P. Seidel. Towards stable and salient multi-view representation of 3d shapes. In M. Spagnuolo, A. Belyaev, and H. Suzuki, eds., *IEEE International Conference on Shape Modeling and Applications 2006 (SMI 2006)*, Matsushima, JAPAN, June 2006, pp. 265–270. IEEE.

15.5 Free-Form Surfaces and Visualization

Coordinator: Christian Rössl

Best-known examples of classical free-form surfaces in computer graphics and animation include tensor-product splines and subdivision surfaces. Immediately related to the particular mathematical model is the process of designing surfaces resulting in the minimization of certain bending energies. Such energy minimization can be used either for fair surface design or for surface deformation. We continue work on various surface editing scenarios and propose new methods. Many data sets are not given explicitly as surfaces but as discrete volume data. In order to visualize such data appropriate continuous models of the data are required. Our work on trivariate splines and here quadratic super splines in particular follows this direction and provides deep analysis of the model as well as new methods for visualization. Furthermore, we investigated on efficient hardware-accelerated rendering of trivariate polynomials and new rendering primitives for scientific visualization.

We remark that we not only developed new methods in this field of surface representation of visualization but also realized relations to other fields in computer graphics such as geometry processing in general, motion capturing, scientific visualization.

15.5.1 Subdivision

Investigators: Ioannis Ivrissimtzis and Rhaleb Zayer

The representation of smooth surface as the limit of an iterative subdivision process applied to polygonal surfaces is a well studied area in computer graphics. We focus on the spectral analysis of the initial input mesh and investigate the relation between polygonal decomposition of 1-ring neighborhoods of a quadrilateral mesh and the eigenvectors of matrices with circulant blocks [1].

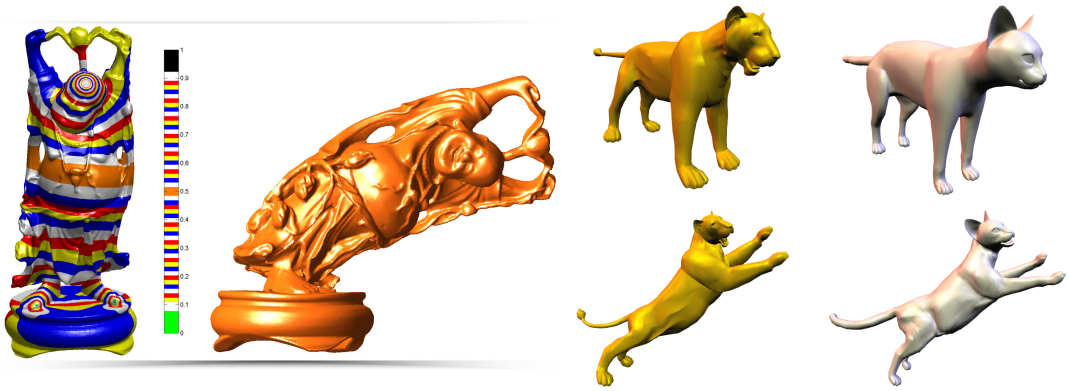


Figure 15.4: The pair of images to the left show how the harmonic field on the Happy Buddha model guides the bending and twisting edit. The image set to the right depicts the deformation transfer of the tiger pose to the cat.

References

- [1] I. Ivrissimtzis, R. Zayer, and H.-P. Seidel. Polygonal decompositions of quadrilateral subdivision meshes. *Computer Graphics & Geometry*, 7(1):16–30, 2005.

15.5.2 Shape Editing

Investigators: Zachi Karni, Christian Rössl, Carsten Stoll, and Rhaleb Zayer

In recent years, interactive shape deformation and editing has been a very active field of research. The goal is the development of algorithms capable of deforming shapes in an intuitive and natural looking way while satisfying a given set of constraints. We investigated several methods based on the discretization of stretching and bending energies using discrete Laplacian operators.

Probably the best-known example of surface editing is when a user selects and drags a subset of vertices to a new location in space, and within a given region of influence the shape follows this manipulation in a natural way. For addressing the challenges raised by the mesh deformation problem, we identify the propagation of local transformations over the surface as a key technique. In order to guide this process, we establish appropriate smooth harmonic scalar fields over the surface which serve as building block for deformation propagation as depicted in Figure15.4.

In addition, we consider a different surface editing scenario: deformation transfer. Here, several manipulations of a reference model are transferred and applied on a target model. A typical application is copying the animation of one object, a potentially costly and time consuming artistic effort, to another similar object automatically and efficiently. As a result, the target object mimics the global deformations of the reference object while its characteristic geometric details are preserved as shown in Figure15.4. In order to achieve such deformation transfer, we apply harmonic fields as a basic tool for efficiently finding correspondence [3].

In interactive shape editing, it is often not necessary to perform large deformations in a single step, as constraints are moved across the screen in small steps each frame. Furthermore, large scale non-linear deformations can be approximated by a series of smaller

Figure 15.5: Example of a deformation of a scanned model based on a set of markers from an optical MoCap system.

linear deformations. Thus we investigated a shape deformation strategy which updates the deformation in real-time by repeatedly solving simple linear equations. This way we can create plausible non-linear deformations at interactive frame rates while guaranteeing high visual quality. As this is only possible for moderately sized models we further enhanced our approach by developing an efficient solution for transferring deformations across different levels of detail, allowing for the editing of high resolution models as well.

Our work on deformation was further tuned for building fast and simple frameworks capable of generating high-quality animations of scanned human characters from input motion data [2, 1]. Additional details are available in Section 15.9.

References

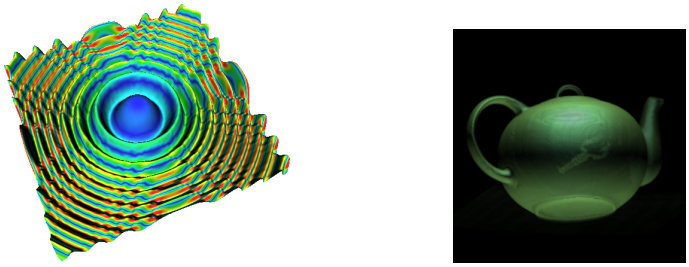
- [1] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Rapid animation of laser-scanned humans. In *IEEE Virtual Reality 2007*, Charlotte, USA, 2007, pp. 223–226. IEEE.
- [2] E. de Aguiar, R. Zayer, C. Theobalt, M. Magnor, and H.-P. Seidel. A framework for natural animation of digitized models. Research Report MPI-I-2006-4-003, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, July 2006.
- [3] R. Zayer, C. Rössl, Z. Karni, and H.-P. Seidel. Harmonic guidance for surface deformation. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, 2005, *Computer Graphics Forum*, vol. 24, pp. 601–609. Blackwell.

15.5.3 Trivariate Splines for Efficient Visualization

Investigator: Christian Rössl

Splines are a well-established and important data model in Computer Aided Geometric Design, probably best known as representation of curves and surfaces. Their application to volume data was only rarely exploited so far. We proposed a new approach to reconstruct non-discrete models from gridded volume samples, e.g., data stemming from a CT or MRI sensor. As a model, we use quadratic trivariate super splines which satisfy appropriate smoothness conditions using polynomial pieces of the lowest possible total degree two. We continued work on trivariate spline models for efficient reconstruction and visualization of volume data:

We studied the smoothness and approximation properties of the quasi-interpolating quadratic super splines in [1]. Here, we observe as a non-standard phenomenon that the derivatives



(a) (b)

Figure 15.6: Volume visualization from qudaratic super spline model: (a) reconstruction of highly-oscillating test data confirms approximation and smoothness properties; (b) shear-warp visualization of the Utah teapot volume data set.

of our splines yield optimal approximation order for smooth data, while the theoretical error of the values is nearly optimal due to the carefully chosen averaging rules. The optimal approximation properties of the derivatives allow to simply sample the necessary gradients directly from the polynomial pieces of the splines for efficient high-quality visualization of iso-surfaces.

When initially proposing the quadratic super spline model, we used ray-casting for rendering iso-surfaces. We showed that for ray-casting intersections with the iso-surface can be computed exactly and efficiently. However, the model is not limited to this technique. Consequently, we showed that alternative volume visualization methods can benefit from quadratic super splines. In particular, we investigated on a combination of our splines with shear-warp techniques and wavelet encoding [2] for efficient high-quality visualization.

References

- [1] G. Nürnberger, C. Rössl, F. Zeilfelder, and H.-P. Seidel. Quasi-interpolation by quadratic piecewise polynomials in three variables. *Computer Aided Geometric Design*, 22:221–249, 2005.
- [2] G. Schlosser, J. Hesser, F. Zeilfelder, C. Rössl, R. Männer, G. Nürnberger, and H.-P. Seidel. Fast visualization by shear-warp on quadratic super-spline models using wavelet data decompositions. In C. T. Silva, E. Gröller, and H. Rushmeier, eds., *16th IEEE Visualization Conference (VIS 2005)*, Minneapolis, MN, USA, 2005, pp. 351–358. IEEE.

15.5.4 Efficient Visualization of Higher Order Surfaces

Investigator: Carsten Stoll

The standard rendering primitive on modern computer hardware is the triangle mesh, a piecewise linear representation of a 3d shape. While triangles are the fastest primitive to render on modern graphics hardware, they suffer from some inherent problems, for example the first order partial derivatives being not smooth, which can lead to noticeable artifacts when the mesh resolution is insufficient. Many recent surface reconstruction algorithms generate surfaces using higher order primitives due to their favorable properties concerning surface

(a) (b) (c)

Figure 15.7: (a) GPU based ray-casting of a SLIM surface with three types of surfaces blended. Rendering performance of roughly 40 frames second is achieved. (b), (c) Using our stylized line primitives for molecule visualization and vector-field streamline visualization.

fitting, like MPU, SLIM, dynamic skin surfaces and higher order iso-surfaces. This, together with the recent advances in GPU hardware capabilities, lead to investigating possibilities to directly render higher order surfaces on the GPU.

We developed an optimized GPU based ray-casting algorithm which allows for the direct visualization of second order primitives and investigated efficient techniques for culling and intersecting [2]. We used our algorithm to directly visualize high-quality high order surfaces generated from SLIM surface reconstruction, quadratic iso-surface in tetrahedral meshes and bilinear quadrilaterals. Figure 15.7 (a) shows an example. Compared to triangle based surface approximations of similar geometric error we achieve only slightly lower frame rates but with much higher visual quality due to the quadratic approximation power of the underlying surfaces.

We also investigated the efficient visualization of a subset of generalized cylinders on the GPU [1]. The stylized line primitives are very useful for scientific visualization, in particular 3d vector-field visualization. Our implementation allows for the efficient rendering of line segments with different visual attributes including color, width, texture, orientation and halo-like effects for improved depth perception and showed the suitability for vector-field visualization (see Section 15.6). Figure 15.7 (b) and (c) show examples.

References

- [1] C. Stoll, S. Gumhold, and H.-P. Seidel. Visualization with stylized line primitives. In C. T. Silva, E. Gröller, and H. Rushmeier, eds., *IEEE Visualization 2005 (VIS 2005)*, Minneapolis, USA, 2005, pp. 695–702. IEEE.

- [2] C. Stoll, S. Gumhold, and H.-P. Seidel. Incremental raycasting of piecewise quadratic surfaces on the GPU. In I. Wald and S. G. Parke, eds., *IEEE Symposium on Interactive Raytracing 2006 Proceedings*, Salt Lake City, USA, September 2006, IEEE Symposium on Interactive Raytracing Proceedings, pp. 141–150. IEEE.

15.6 Vector Field Visualization and Applications of Vector Field Techniques

Coordinator: Holger Theisel

The main focus of the group was to do research on visualization techniques for vector fields. There, the focus was on topological methods and on the extraction of vortex core structures. Topological methods aim in segmenting the flow into areas of similar flow behavior. Therefore they offer to represent even complex flow structures with only a limited number of graphical primitives. Vortex core structures are important flow features which are essential for the understanding of several flow phenomena. In addition to this research, the group has worked on a number of different problems in the field of visualization and vector field processing.

15.6.1 Vector Field Based Shape Deformations

Investigators: Wolfram von Funck and Holger Theisel

Shape deformations is a well-researched area in computer graphics and animation with many applications ranging from automotive design to movie production. Recent approaches emphasize intuitive and physically plausible editing while maintaining interactive performance. Volume-preservation and prevention of self-intersections of the deforming shape are important aspects when natural and realistic deformations are desired. While these constraints usually require complex computations, we address these issues by our vector field based approach [1]. Our method defines shape deformations by constructing and interactively modifying C^1 continuous time-dependent divergence-free vector fields. The deformation is obtained by a path line integration of vertices. Figure 15.8 illustrates this idea. This way, the deformation is volume-preserving, free of (local and global) self-intersections, feature preserving, smoothness preserving, local, and independent of the underlying shape

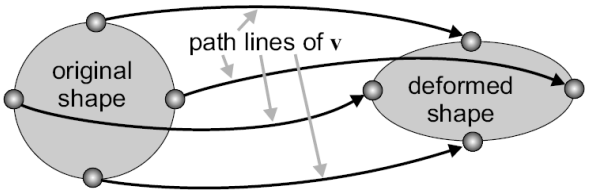


Figure 15.8: Vector field based shape deformation: every vertex of the original shape undergoes a path line integration of a divergence-free vector field $\mathbf{v}$ to find its new position.

Different modeling metaphors support the approach which is able to modify the vector field on-the-fly according to the user input. Being independent of the shape representation, a GPU implementation of the approach is able to achieve interactive frame rates for moderate mesh sizes, and the numerical integration preserves the volume with a high accuracy.

Figure 15.9 shows a possible application scenario, where a hand model which is manipulated using our approach. The deformation looks realistic and giveS the impression of real skin and tissue of the hand, even though no physical or anatomical model is involved.

Figure 15.9: A hand model is deformed in a volume-preserving manner without self-intersections. Even though no skeletal hand model or physical simulation is involved, the result looks natural and physically plausible.

15.6.2 Multifield Visualization

Investigators: Natascha Sauber and Holger Theisel

Most of the simulations in computer science and engineering create as output not only one, but multiple fields describing different aspects of the simulated topic. These fields, that can be scalar, vector or tensor fields, are usually defined over the same 3D domain. There are cases when it is insufficient or inconvenient to visualize the fields separately. Some fields may be redundant, while others may have an important spatial or temporal relation to each other which is worth visualizing. There are many established techniques for visualizing a single scalar field, like slicing, isosurfacing, or direct volume rendering. However, the simultaneous visualization of multiple fields, or the visualization of relation between them is still a challenge.

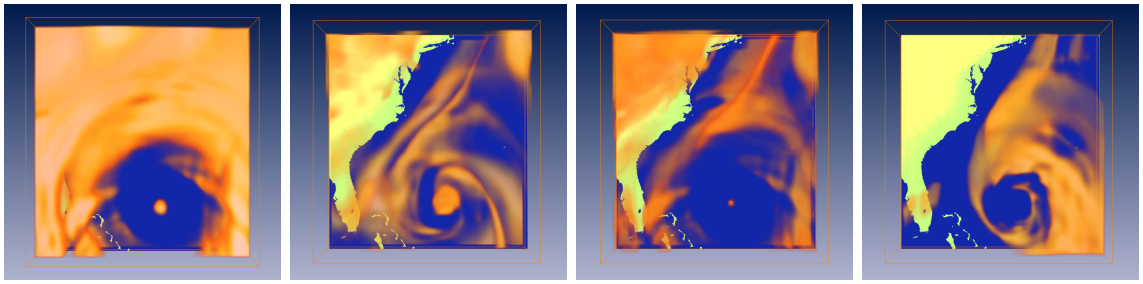
We present an approach to visualizing correlations in 3D multifield scalar data [2]. The core of our approach is the computation of correlation fields, which are scalar fields containing the local correlations of subsets of the multiple fields, and thus relational information. While the visualization of the correlation fields can be done using standard 3D volume visualization techniques, their huge number makes selection and handling difficult. We introduce the Multifield-Graph to give an overview of which subsets of the multiple fields correlate and to show the strength of their correlation. This information guides the selection of informative correlation fields for visualization. We use our approach to visually analyze a number of real and synthetic multifield datasets. One of the multifield datasets we analyzed is the simulation of a hurricane. It consists of 6 scalar fields representing physical properties of the hurricane, like pressure, vapor, and temperature. Some of the resulting correlation fields can be seen in Figure 15.10.

15.6.3 Path Line Oriented Flow Topology on Time-dependent Flow Fields

Investigators: Kuangyu Shi and Holger Theisel

In this project we introduce an approach to extracting a path line oriented topological segmentation for periodic 2D time-dependent vector fields [4].

For time-dependent vector fields there exists a number of relevant characteristic curves, such as stream lines, path lines, streak lines and time lines. Among them, stream lines and path lines have the uniqueness property: through each point in the space-time domain there



$\mathbf{C}_{\{3,5\}}$ $\mathbf{C}_{\{4,5\}}$ $\mathbf{C}_{\{3,4,5\}}$ $\mathbf{C}_{\{0,2,4\}}$

Figure 15.10: Visualization of local correlation between multiple scalar fields of a hurricane simulation. The sub-figures labeled with $\mathbf{C}_{\{i,j,k\}}$ show volume visualizations of the local correlation between the scalar fields i, j, k, while field 0 contains cloudiness, field 3 pressure, field 4 vapor, and field 5 temperature. (data set by the Weather Research and Forecast (WRF) model, courtesy of NCAR and the U.S. National Science Foundation (NSF))

is exactly one stream line and one path line passing through. This gives that two different kinds of topologies can be considered: a stream line oriented topology segmenting areas of similar stream line behavior, and a path line oriented topology which does so for path lines.

Topological methods aiming in capturing the asymptotic behavior of path lines rarely exist because path lines are usually only defined over a fixed time-interval, making statements about their asymptotic behavior impossible. For the data class of periodic vector fields, this restriction does not apply any more. Through analyzing a 2D Poincaré map, we use a point based method to classify the asymptotic behavior for each point.

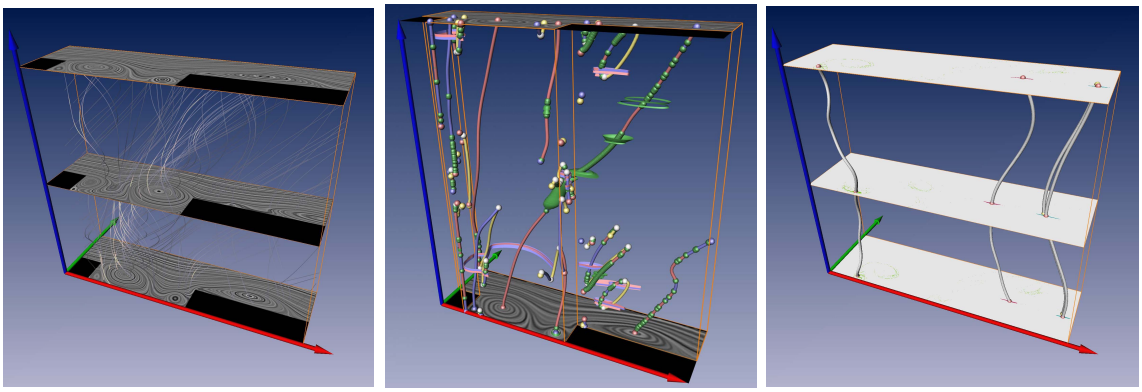


Figure 15.11: The left shows a visualization of the path lines in a cavity data set; The middle illustrates a stream line oriented topology visualization for the same data set; The left shows the result of our approach on cavity data set, a path line oriented visualization is presented with critical path lines and basins for forward integration. (data set by Mo Samimy and Edgar Caraballo (both Ohio State University) as well as Bernd R. Noack (TU Berlin))

We made the following contributions:

- We introduced an approach to analyzing the asymptotic behavior of path lines in periodic time-dependent vector fields.
- We defined, extracted, and classified critical path lines.
- We computed the basins from which the path lines converge to the critical path lines in forward or backward integration.

Our examples show that the path line oriented topology gives significantly different topological information than the stream line oriented one. Figure 15.11 is a visualization of a vector field describing the flow at a 2D cavity

15.6.4 Segmentation of DT-MRI Anisotropy Isosurfaces

Investigators: Thomas Schultz and Holger Theisel

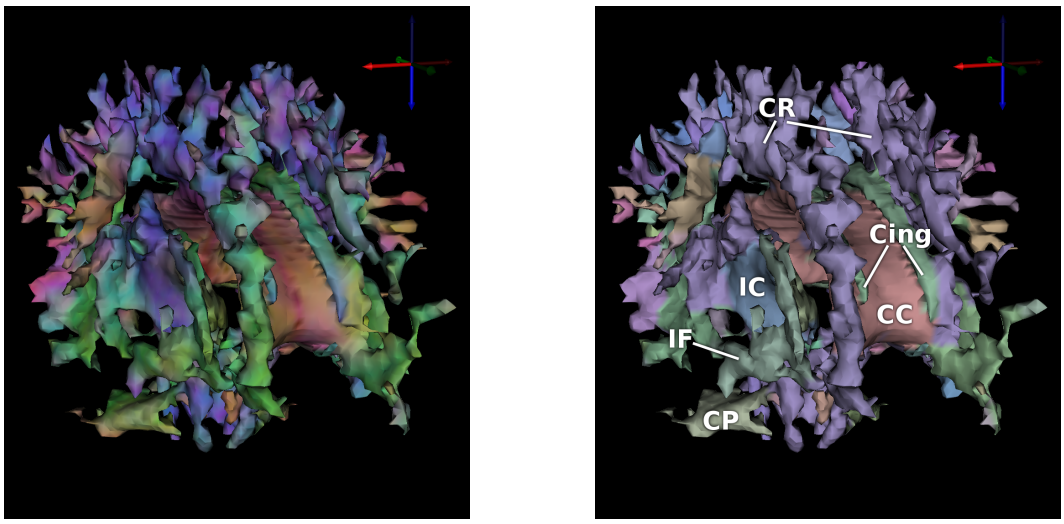
Diffusion-tensor magnetic resonance imaging (DT-MRI) is a medical imaging modality that measures the self-diffusion of water molecules. In each voxel, the method determines a symmetric 3×3 matrix (a *diffusion tensor*) that models the observed distribution of apparent diffusivities. When applied to the human brain, DT-MRI offers the unique possibility to examine the orientation of major nerve fiber tracts.

Anisotropy measures reduce the tensor data to a scalar index that reflects the degree to which the apparent diffusivity in a voxel is directionally dependent. In the human brain, high anisotropy indicates coherently organized nerve fibers, so isosurfaces along which the anisotropy is constant outline the contours of major fiber tracts. Thus, anisotropy isosurfaces give a large-scale overview of the data, which allows to identify important anatomic structures.

Our contribution [3] is to suggest a method that uses the part of the tensor information which has been neglected by the anisotropy metric to partition anisotropy isosurfaces into regions that correspond to anatomic units revealed by the data. For the implementation, we propose an edge-based watershed method which adapts and significantly extends methods that have previously been used in the context of curvature-based mesh segmentation.

Figure 15.12(a) shows a sample anisotropy isosurface, colored in the XYZ-RGB color scheme (red for "left-right", green for "front-back", blue for "top-bottom"). Figure 15.12(b) presents a segmentation result, which assigns a uniform color to each region, reflecting the average diffusion tensor from the respective region. Note that the surface is a single connected component, so all shown regions are defined by our algorithm.

The annotations in Figure 15.12(b) illustrate that our method successfully segmented a number of anatomic structures: The *corpus callosum* (CC) is separated from the *cingulum bundle* (Cing), and from the *corona radiata* (CR). Moreover, regions have been found that correspond to the *internal capsule* (IC), the *cerebellar peduncle* (CP), and the *inferior fronto-occipital fasciculus* (IF). A possible application of our method is interactive clipping of regions, which facilitates examination of the complex and convoluted surfaces.



(a) (b)

Figure 15.12: While sub-figure (a) shows an anisotropy isosurface of the human brain using a standard color scheme, sub-figure (b) presents an annotated segmentation result from our algorithm. (Dataset provided by Alfred Anwander, MPI Human Cognitive and Brain Science, Leipzig, Germany)

15.6.5 Parallel Vector Surfaces in 3D Time-dependent Flow Fields

Investigator: Holger Theisel

We introduce an approach to tracking vortex core lines in time-dependent 3D flow fields [5] which are defined by the parallel vectors approach. They build surface structures in the 4D space-time domain. To extract them, we introduce two 4D vector fields which act as feature flow fields, i.e., their integration gives the vortex core structures. This way, the extraction/tracking of vortex core lines is reduced to a simple stream line/surface integration of vector fields. We choose this approach because of the following reasons:

- Numerical stream line/surface integration is well-understood in the Visualization community. A variety of fast and stable algorithms exist for this purpose.
- The stream surface integration approach is independent of an underlying grid, giving a subcell accuracy and relieving us of finding appropriate local connection strategies.
- Bifurcations (i.e. events of sudden changes of the behavior of vortex core lines over time) play an important role in the understanding of the dynamical behavior of vortex core lines. Contrary to pre-existing methods, the FFF approach permits to localize, characterize and classify these bifurcations.

As part of this approach, we extract and classify local bifurcations of vortex core lines in space-time. Based on a 4D stream surface integration, we provide an algorithm to extract

the complete vortex core structure. Figure 15.13 gives an illustration.

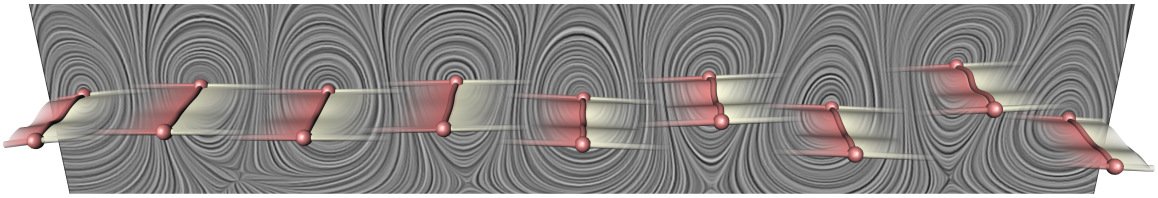


Figure 15.13: Flow behind a circular cylinder. Shown are vortex core lines in a certain frame of reference. Their evolution over time is tracked by our algorithm and depicted using transparent surfaces. Red color encodes the past while gray shows the future. (data set by Bernd R. Noack (TU Berlin) from a direct numerical Navier Stokes simulation by Gerd Mutschke (FZ Rossendorf))

References

- [1] W. von Funck, H. Theisel, and H.-P. Seidel. Vector field based shape deformations. In J. Dorsey, ed., *Proceedings of ACM SIGGRAPH 2006*, Boston, MA, USA, July 2006, *ACM Transactions on Graphics*, vol. 25, pp. 1118–1125. ACM. Proc. of ACM SIGGRAPH '06.
- [2] N. Sauber, H. Theisel, and H.-P. Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. In E. Gröller, A. Pang, C. T. Silva, J. Stasko, and J. van Wijk, eds., *IEEE Visualization Conference 2006*, Baltimore, USA, November 2006, *IEEE Visualization*, vol. 12, pp. 917–924. IEEE.
- [3] T. Schultz, H. Theisel, and H.-P. Seidel. Segmentation of DT-MRI anisotropy isosurfaces. In K. Museth, T. Möller, and A. Ynnerman, eds., *Proc. Eurographics / IEEE-VGTC Symposium on Visualization (EuroVis '07)*, Norrköping, Sweden, May 2007. Eurographics. Accepted for publication.
- [4] K. Shi, H. Theisel, T. Weinkauf, H. Hauser, H.-C. Hege, and H.-P. Seidel. Path line oriented topology for periodic 2D time-dependent vector fields. In B. Sousa Santos, T. Ertl, K. I. Joy, D. W. Fellner, T. Möller, and S. N. Spencer, eds., *EUROVIS 2006, Eurographics / IEEE VGTC Symposium on Visualization*, Lisbon, Portugal, 2006, pp. 139–146. Eurographics.
- [5] H. Theisel, J. Sahner, T. Weinkauf, H.-C. Hege, and H.-P. Seidel. Extraction of parallel vector surfaces in 3d time-dependent fields and applications to vortex core line tracking. In C. T. Silva, E. Gröller, and H. Rushmeier, eds., *IEEE Visualization 2005 (VIS 2005)*, Minneapolis, USA, 2005, pp. 631–638. IEEE.

15.7 Modeling and Animation of Faces and Hands

Coordinator: Volker Blanz

The goals of our learning-based approach to Computer Graphics are to enhance the realism and to increase the complexity of virtual objects, but at the same time to reduce the manual work involved in the production of 3D content. Learning-based methods are becoming a major trend in Graphics today, and our group has played an active role in initiating and developing this technique.

Learning-based graphics involves three main steps: (1) capturing data from real-world objects, such as 3D scans of human faces, (2) statistical analysis of the data, and (3) application to new, real-world data, such as scans or images. The method retains the appearance and complexity of real objects, but allows the user to edit them in an efficient and meaningful way. On the following pages, we describe a number of projects that implement this strategy on 3D scans of human faces and hands.

In a collaboration with MPI for Biological Cybernetics, we have developed an algorithm that uses state-of-the-art machine learning for processing raw 3D scan data [9, 8]. In this work, we have modeled the point clouds of 3D scan data as implicit surfaces, using Support-Vector Regression, and used this implicit representation to establish point-to-point correspondence between pairs of scans using another Support Vector Regression. The applications of this algorithm are in data post-processing, hole-filling, smoothing and 3D registration of scans.

In a joint project with University of Dallas, Texas, we have used the statistical model of 3D faces to investigate the mechanisms of human perception in a series of psychological experiments [3, 4]. As stimuli, we used morphs between original 3D scans of male faces, and the average male face. By morphing faces beyond the average, we created anti-faces [5], in which all facial features are inverted within the range of appearances found in a natural population: For example, if an original face has dark eyebrows, the anti-face will have bright eyebrows, and if the original face is skinny, the anti-face will look obese. We have established previously [5] a perceptual aftereffect in the perception of human faces: If a participant is shown an anti-face and subsequently the average face, they will briefly perceive the original face. Unlike the well-known aftereffects in color vision, this is not an after-image close to the retina level, but an adaptation effect in high-level processing units of the brain. In recent work, we have shown that this aftereffect can be found even across changes in 3D pose of the stimulus faces, and that it involves both shape and texture [3]. These findings have direct implications on the models of neural representations of faces, indicating a prototype-related representation on the level of viewpoint-independent processing, and a fusion of shape and texture information on that level.

The 3D morphing technique in our statistical face model has also been used in a prototype application for forensic purposes [2]. In our software system, the user can create a composite face of a person based on a number of different high-level user interfaces. We have evaluated the system with a forensic artist in a test scenario where witnesses had to reconstruct the faces of suspects. Our system is more efficient and intuitive than existing software, and it may help to simplify the tedious and often painful process of making composite faces for crime witnesses or victims.

15.7.1 Gesture Modeling and Animation

Investigators: Michael Neff and Irene Albrecht

Animated characters that move and gesticulate appropriately with spoken text are useful in a wide range of applications. Unfortunately, this class of movement is very difficult to generate, even more so when a unique, individual movement style is required. In [1, 6], we present a system that is capable of producing gesture animation for given input text in the

style of a particular performer. While we concentrate on hand-arm gestures, the approach is also applicable to the entire body.

Our process starts with video of a human performer whose gesturing style we wish to animate. A tool-assisted annotation process is first performed on the video, which serves to build a statistical model of the person's particular gesturing style. Using this model and input text with information structure tags, our generation algorithm creates a gesture script. As opposed to isolated singleton gestures, this gesture script specifies a stream of continuous gestures that are coordinated with speech and naturally merge into each other. The script is then passed to an animation system, which enhances the gesture description with additional detail, such as augmented timing information and character specific pose variation. An animation sub-engine then generates either kinematic or physically simulated motion based on this description.

In total, the system is capable of creating animation that replicates a particular performance in the video corpus, and of generating performances of novel text that are consistent with a given performer's style.

As a person changes the orientation of his relaxed hand, finger and wrist angles will change passively due to the influence of gravity. Modeling this effect adds important subtlety to character motion. We present a technique for modeling these deformations [7]. A dynamic model of the hand is built using Proportional-Derivative controllers as a first order approximation to muscles. A process for tuning the model to match the relaxed hand shape of subjects is discussed. Once the model is tuned, it can be used to sample the space of all possible arm orientations and samples of wrist and finger angles are taken. From these samples, a kinematic model of passive hand deformation is built. Either the tuned dynamic model or the kinematic model can be used to generate final animations. These techniques increase the realism of gesture animation, where the character often maintains a relaxed hand.

References

- [1] I. Albrecht, M. Kipp, M. Neff, and H.-P. Seidel. Gesture modeling and animation by imitation. Research Report MPI-I-2006-4-008, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, August 2006.
- [2] V. Blanz, I. Albrecht, J. Haber, and H.-P. Seidel. Creating face models from vague mental images. In L. Szirmay-Kalos and E. Gröller, eds., *EUROGRAPHICS 2006 (EG'06)*, Vienna, Austria, September 2006, *Computer Graphics Forum*, vol. 25, pp. 645–654. Blackwell.
- [3] F. Jiang, V. Blanz, and A. J. O'Toole. Probing the visual representation of faces with adaption. a view from the other side of the mean. *Psychological Science*, 17(6):493–500, 2006.
- [4] F. Jiang, V. Blanz, and A. J. O'Toole. The role of familiarity in three-dimensional view transferability of face identity adaptation. *Vision Research*, to appear, 2007.

[5] D. A. Leopold, A. J. O'Toole, T. Vetter, and V. Blanz. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, 4(1):89–94, 2001.

- [6] M. Neef, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics*, 26, 2007.
- [7] M. Neff and H.-P. Seidel. Modeling relaxed hand shape for character animation. In F. J. Perales and R. B. Fisher, eds., *Articulated motion and deformable objects, 4th International Conference, AMDO 2006*, Port d'Andratx, Spain, 2006, *LNCS 4069*, pp. 262–270. Springer.

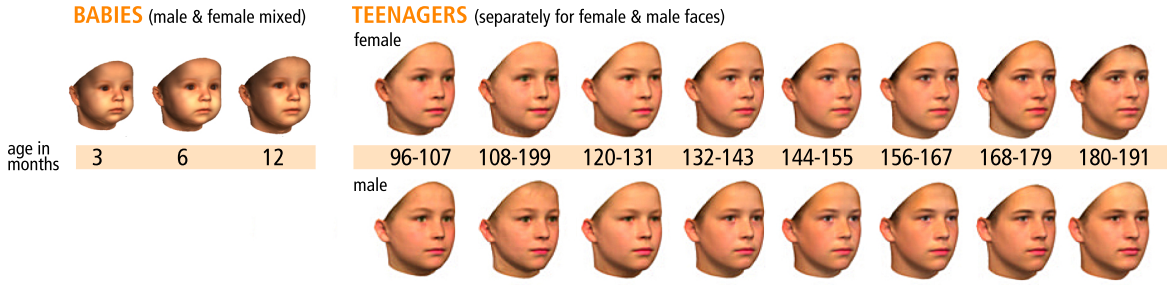


Figure 15.14: Averages of age groups for female and male faces. Each age group comprises an age interval of 12 months. In the age group 3-12 months we did not differentiate between female and male faces

- [8] B. Schölkopf, V. Blanz, and F. Steinke. Object correspondence as a machine learning problem. In *International Conference on Machine Learning ICML*, Bonn, Germany, 2005, pp. 776–783. ACM Press.

- [9] F. Steinke, B. Schölkopf, and V. Blanz. Support vector machines for 3d shape processing. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, 2005, *Computer Graphics Forum*, vol. 24, pp. 285–294. Blackwell.

15.7.2 Learning-Based Object Modelling

Investigators: Volker Blanz, Kristina Scherbaum, and Robert Bargmann

Prediction of Facial Growth

Police investigators searching for children that have been missing for several years face the problem to predict the children's current looks from images taken at an earlier age. Today, much of this work is done manually: given an old image of the missing child, experienced and skilled forensic artists estimate his current look by retouching and redrawing his face shape in the image by hand. In order to simplify and automate this task, we present a method that learns from a large dataset of 3D scans of faces how children grow, and applies this transformation to 3D faces and to images.

To collect a sufficient amount of 3D face data we acquired 250 facial 3D scans of children and teenagers in schools. Supplementary the Universitätsklinikum Tübingen provided a dataset of 3D laser scans of babies, as a result of collaboration. By fitting a 3D Morphable Model of faces ([2]) to each of these 3D scans we established full correspondence over all faces in 3D space and represented them as vectors in a high-dimensional face space. As a result of a statistical analysis and a division into age groups we could compute a set of age-based average faces (Figure 15.14).

The size of the age groups is a tradeoff between temporal resolution of the growth simulation, and statistical significance of the changes observed. Concerning our dataset of baby faces the age groups were chosen according to the given scanning intervals (3, 6 and 12 months). For the babies, each group consists of both, boys and girls since no remarkable

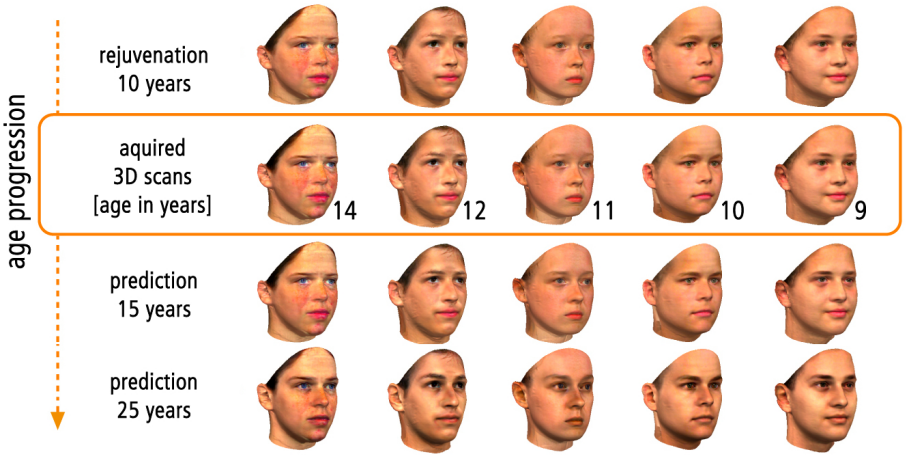


Figure 15.15: The figure shows example faces of the face database. Each face is transformed into several ages by applying our piecewise linear growth function. Individual characteristics of the faces are retained by the age transformation.

physiognomic differences appear at that age. For the teenager faces, we achieved best results with age intervals of 1 year and with separate groups for girls and boys.

To simulate aging of a particular 3D face, a piecewise linear approach shifts the face along the directions that connect the averages of different age groups (Figure 15.15). When the age-transformation is applied over a larger age period, the age-specific average head grows in size, while the individual difference vector does not. As a consequence, the relative displacement, and thus the relative distinctiveness, is diminished. For example, a nose that is 10 percent larger than the average at an early age would be only 5 percent larger in the age-transformed face. We evaluated how much the faces grow among teenagers, using a size measure to compute an overall scaling factor.

For aging simulation in images, a 3D model of the initial face is reconstructed and after applying an age-transformation on shape and texture, the face is rendered back into images of children. By rendering the age-transformed face into images of children with different hairstyles, our system generates a variety of possible appearances. In the best case, the prediction made by our algorithm can only give the most likely image of what a child may look like in the future. Our approach does not incorporate environmental factors, such as nutrition, exposure to sunlight or physical activity.

In addition to the generation of photofit pictures of missing children our system may be used in the fields of character design, animation and morphing for special effects.

Learning Articulation

In this project ([1]), we studied the natural articulation of the mouth from a real person so to be able to resynthesize novel face animation from any novel audio file. This investigation aims at letting the mouth movements of virtual speakers appear more realistic and at the same time at replacing gradual animation method for studio animations by automatic procedures so that in the future, movie companies will focus more on computer animation and the

artistic arrangement of the movements when speaking and create thus more expressive, more realistic and more subtle contents.

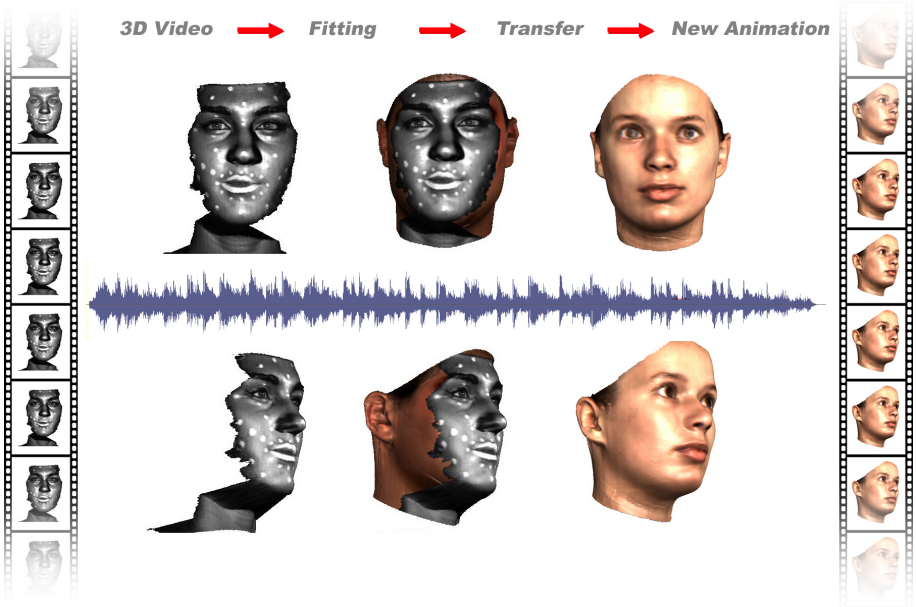


Figure 15.16: **Learning Articulation:** To build the system, we go through several steps: first, 3D video is recorded from a real speaking person; a morphable model is then fitted to every 3D face to complete the meshes to full featured faces; the data is then registered so that we were able to statistically analyze the face deformation over time; in further step, we are able to synthesize face animation from novel audio sequences and to transfer this articulation to different faces.

The data is acquired through a high speed dynamic structured-light based 3D scanner. A real person utters a set of sentences while the surface of its face is recorded in 3D with a frame-rate of 40 frames per second. Alongside, the audio is also recorded to get a correspondence between the frames and the uttered phonemes.

The scanner's output is a stream of independent 3D meshes over the time. Moreover, the scanner recreates only partially the surface of the face as it cannot reconstruct hidden or shadowed regions. To remedy this, the surface is preprocessed: holes are filled, surfaces is smoothened and furthermore, the lateral parts of the head are reconstructed by fitting a morphable model to each 3D frame. The data has then to be registered as rather than having a stream of independent 3D meshes we want to transform it into a mesh that deforms over the time. All meshes are thus put into correspondence with a selected one by mean of an optical flow algorithm.

The deformation of the mesh is then statistically analyzed by performing a modified principal component analysis (PCA): the PCA will focus on the deformation of the mouth without being influenced by the blinking of the eyes. The PCA will describe a viseme-space, in which all recorded 3D frames can be projected onto. Each frame being labeled with its corresponding phoneme, we are able to define viseme clusters inside this space where any

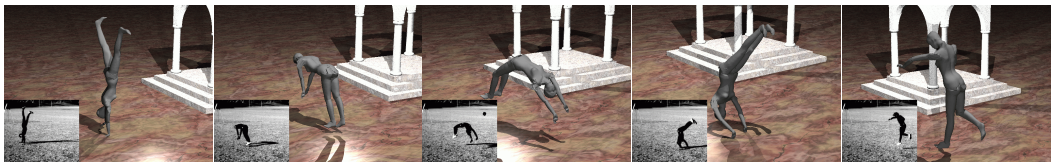


Figure 15.17: Markerless motion capture: Reanimated Cartwheel-Flick-Flack sequence in a virtual environment. The small images show one of the four used camera views.

curve describes a new articulation.

Once we know the position of the different clusters, a novel articulation can be synthesized from an audio file by simply connecting the clusters in the same sequence as the phonemes from the novel sentences, and interpolate between them along the time to generate a natural novel 3D articulation.

In a last step, we can put novel faces in correspondence with our model, and thus directly transfer the articulation to any new face.

References

- [1] R. Bargmann, V. Blanz, and H.-P. Seidel. Learning-based facial rearticulation using streams of 3D scans. In B.-Y. Chen, ed., *The 14th Pacific Conference on Computer Graphics and Applications*, Taipei, Taiwan, October 2006, pp. 232–241. National Taiwan University.

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Computer Graphics Proc. SIGGRAPH'99*, 1999, pp. 187–194.

15.8 Markerless Motion Capture

Coordinator: Bodo Rosenhahn

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 reasons, marker-less tracking is an important field of research that requires knowledge in biomechanics, computer vision and computer graphics.

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) [1, 10]
- 2D-3D pose estimation (based on free-form surface meshes) [15]
- Correspondence estimation [8].
- Particle filter [6, 3]
- Statistical learning [4, 6, 2]
- Quantitative error analysis [9, 14]
- Cloth synthesis and animation [13, 12, 7]
- Texture driven tracking [11, 5]

In the following we will describe some of these projects

15.8.1 Statistical Learning

Investigator: Bodo Rosenhahn

The aim of this project is to supplement of prior knowledge about joint angle configurations in the scope of 3-D human pose tracking. Training samples obtained from an industrial marker based tracking system are used for a non-parametric Parzen density estimation in the high dimensional joint configuration space. The probability densities constrain the image-driven joint angle estimates by drawing solutions towards familiar configurations. This prevents the method from producing unrealistic pose estimates due to unreliable image cues. Experiments on sequences with a human leg model reveal a considerably increased stability, particularly in the presence of disturbed images and occlusions, see Figure 15.18

15.8.2 Textured Model Based Tracking

Investigators: Jürgen Gall and Bodo Rosenhahn

Estimating the pose of a rigid body means to determine the rigid body motion in the 3D space from 2D images. For solving this problem, we exploit the available information on the object, namely shape and texture. Knowing the 3D shape, the estimating process relies on correspondences between some 2D features in the images and their counterparts on the 3D model. In [10], the surface of the model is matched to the contour extracted by a level-set segmentation, see Figure 15.19(b). Although the silhouette provides stable correspondences at the projected border of the object, it does not always contain enough information for convex and symmetric objects to estimate the pose uniquely. Furthermore, the contour extraction limits the approach for rather slow movements since the segmentation gets easily stuck in a local optimum.

We extended the contour based tracker by incorporating the texture of the object [5]. By matching local descriptors of interest points between the textured model projected onto the image plane and the images of a frame, we obtain additional reliable correspondences, see Figure 15.19(a). Since these features are invariant to image rotation and translation, they are suitable for fast movements and for detecting the initial pose automatically. However, they are sometimes not well distributed on the object and the number of matches varies from frame to frame, thereby making the pose estimation difficult. We demonstrated in [5], that the fusion of these complementary features, namely contour and local descriptors, increases

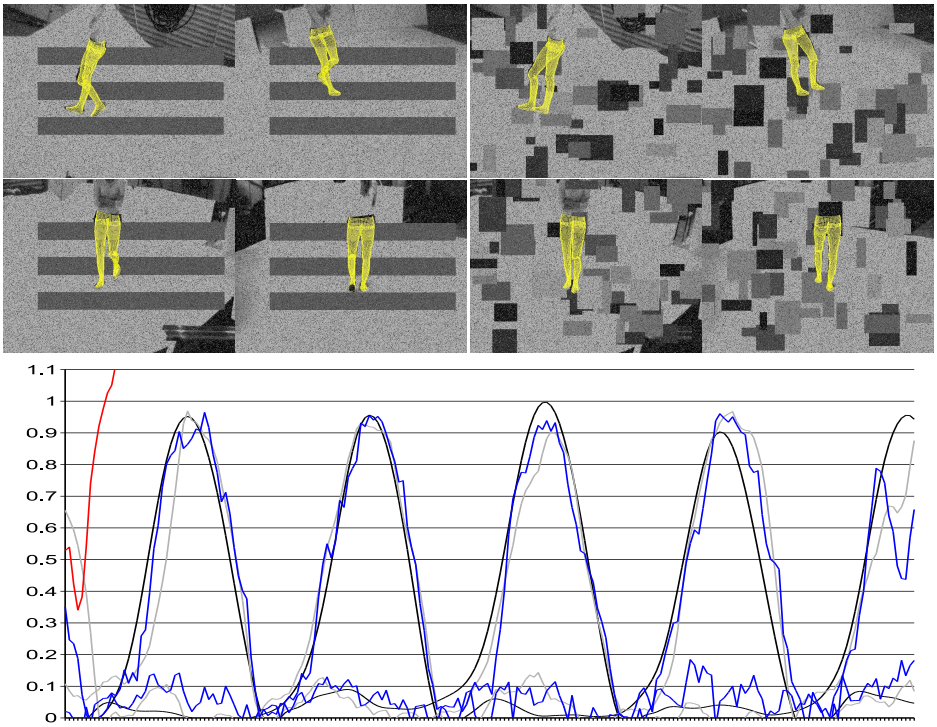
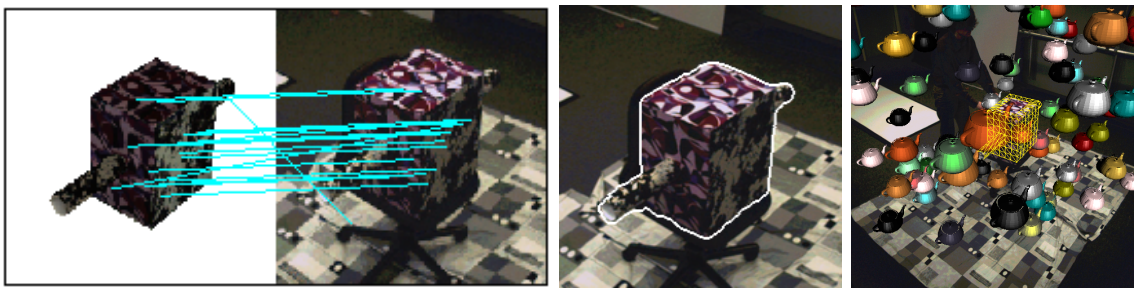


Figure 15.18: Top: Pose estimates of a leg model in a sample disturbed by three enduring gray bars plus 25% uncorrelated pixel noise (top left) or artificial occlusion by random rectangles (top right). Bottom: Joint angles of the left and right knee, respectively. **Black:** marker based system (ground truth). **Gray:** occlusion by permanent bars. **Blue:** occlusion by random rectangles. **Red:** tracking without prior fails after a couple of frames.



(a) (b) (c)

Figure 15.19: *a):* Correspondences extracted from the texture between the textured model and an image. *b):* The extracted silhouette provides complementary correspondences. *c):* The method is robust to occlusions. The object is tracked while 80 teapots are dropping from the sky.

the robustness of the pose estimation, particularly, in the case of occlusions as shown in Figure 15.19(c).

15.8.3 Interacting Particle Systems for Motion Capture

Investigators: Jürgen Gall and Bodo Rosenhahn

Figure 15.20: Particles with a higher weight are brighter, particles with a lower weight are darker. The particles converge to the pose with the lowest energy as the number of steps increases.

Interacting particle systems approximate a distribution of interest by a finite number of random variables, called particles, where the particles interact between the time steps. In computer vision, they are commonly known as particle filters and are used for solving a non-linear and non-Gaussian filtering problem. In the context of human motion capture, it means that the unknown parameters of the human, e.g. joint angles and position of the human body, are estimated from a sequence of observations, namely images. However, these systems also apply for trapping analysis, evolutionary algorithms, statistics, and optimization. Concerning the latter, we derived from the mathematical theory of interacting particle systems a method for global optimization, which we call interacting simulated annealing [3]. We demonstrated in [6] that it can be applied for human motion capture, see Figure 15.20.

In general, it is still an open question whether modelling human motion capturing as optimization problem or as filtering problem is better. The optimization problem is mostly solved by iterative methods like gradient descent approaches. They work very well as long as the starting point is near the global optimum, however, they get easily stuck in a local optimum. The filtering problem is often solved by particle filters. Instead of a single solution, they provide a probability distribution on the set of parameters making them robust to uncertainty. However, finding a good model for the unknown dynamics and the unknown likelihood is very challenging. Our approach overcomes the dilemma of local optima and the poor performance of a generic particle filter in the case of an inexact likelihood function as shown in [3].

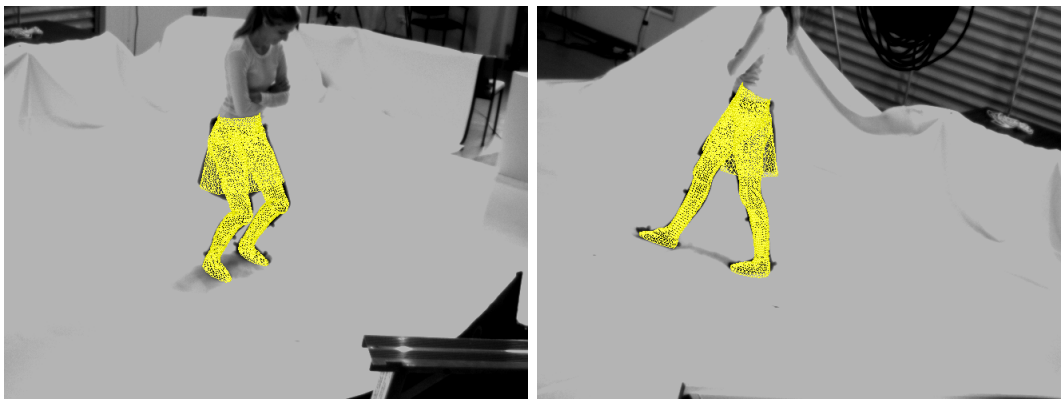


Figure 15.21: Frames from two tracking sequences overlaid with the wireframe representation of the estimated poses.

15.8.4 Cloth Simulation Based Vision

Investigators: Nils Hasler and Bodo Rosenhahn

Cloth simulation has recently been the subject of a substantial amount of research from the computer graphics community. Yet, the primary goal of the investigations was to develop plausible animations rather than realistic simulations. We try to employ cloth simulation to a different class of problems. By reconstructing an observed scene using a cloth simulation we are able to support vision based algorithms in analyzing the scene. Two different applications of the fundamental idea were investigated. On the one hand a silhouette based motion capture algorithm was enhanced with the cloth simulation to allow the tracking of loosely dressed humans. On the other hand 3D-scans were semantically segmented into dressed and unclad parts by fitting the parameterizable simulation model of a garment to the scan data. Following this procedure it was possible to extract the dimensions of the garments as well as segment the scan.

Motion Capture

A state-of-the-art markerless motion capture algorithm was augmented with a cloth simulation to allow tracking of dressed humans. The original approach iteratively runs a level set segmentation algorithm and a twist-based pose estimation procedure to compute the pose of the tracked person. The augmented version additionally runs the cloth simulation every time the underlying mesh of the human model is deformed to obtain the dynamically correct development of the proposed configuration change from one frame to the next. The same optimization procedure could then be used to track a dressed person.

The employed cloth simulation is based on a continuous cloth model which is improved by a bending model and realistic air resistance forces. By running the conducted experiments in parallel with a commercial marker based motion capture system we are able to show that our technique achieves about the same accuracy as the commercial system (approximately 3°).

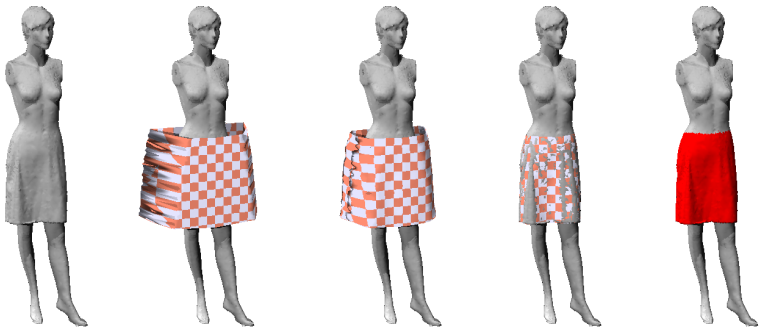


Figure 15.22: Segmentation of 3D-scan data into unclad and garmented parts is achieved by draping a parameterized clothing model on the scan.

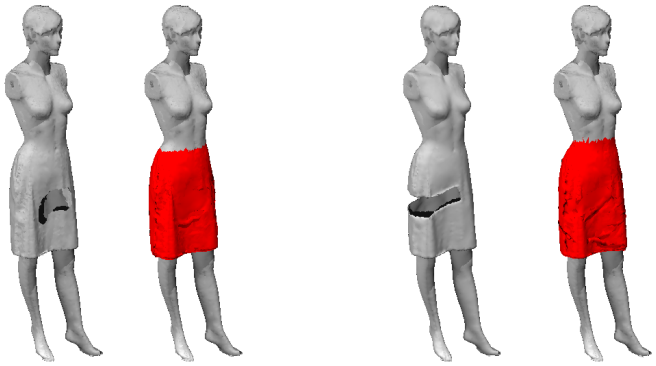


Figure 15.23: The analysis-by-synthesis technique can also be applied to corrupted meshes. Holes are then filled automatically.

Semantic Mesh Segmentation

Investigators: Nils Hasler and Bodo Rosenhahn

An other analysis-by-synthesis application of textile simulations is presented in [7]. Here we apply the algorithm to segmenting 3D-scans of dressed humans. That is parameterizable garment models are draped on 3D laser scans of dressed humans. The general draping procedure is displayed in Figure 15.22. Garments are manually placed roughly at the expected position above the 3D scan. Then the cloth simulation drapes the garment on the scan. A simulated annealing optimization procedure is subsequently able to estimate the dimensions of the attire. The resulting configuration can ultimately be used to segment the scan into dressed and naked parts. An additional interesting feature of the approach is that it can easily handle severely corrupted scan data as shown in Figure 15.23.

References

- [1] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. High accuracy optical flow serves 3-D pose tracking: Exploiting contour and flow based constraints. In A. Leonarids, H. Bishof, and A. Prinz, eds., *Computer vision - ECCV 2006, 9th European Conference on Computer Vision; Part II*, Graz, Austria, 2006, *LNCS 3952*, pp. 98–111. Springer.
- [2] T. Brox, B. Rosenhahn, U. Kersting, and D. Cremers. Nonparametric density estimation for human pose tracking. In K. Franke, K.-R. Müller, B. Nickolay, and R. Schäfer, eds., *Pattern Recognition, 28th DAGM Symposium*, Berlin, Germany, 2006, *LNCS 4174*, pp. 546–555. Springer.
- [3] J. Gall, J. Potthoff, C. Schnoerr, B. Rosenhahn, and H.-P. Seidel. Interacting and annealing particle filters: Mathematics and a recipe for applications. *Journal of Mathematical Imaging and Vision*, X, 2007.
- [4] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Learning for multi-view 3D tracking in the context of particle filters. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, and T. Malzbender, eds., *Advances in Visual Computing, Second International Symposium, ISVC 2006, Part II*, Lake Tahoe, NV, USA, 2006, *LNCS 4292*, pp. 59–69. Springer.
- [5] J. Gall, B. Rosenhahn, and H.-P. Seidel. Robust pose estimation with 3D textured models. In L.-W. Chang and W.-N. Lie, eds., *Advances in Image and Video Technology, First Pacific Rim Symposium, PSIVT 2006*, Hsinchu, Taiwan, December 2006, *LNCS 4319*, pp. 84–95. Springer.
- [6] 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*. Springer, Heidelberg, 2007.
- [7] N. Hasler, B. Rosenhahn, and H.-P. Seidel. Reverse engineering garments. In A. Gagalowicz and W. Philips, eds., *MIRAGE 2007*, Rocquencourt, France, March 2007, *LNCS 4418*, pp. 200–211. Springer.
- [8] B. Rosenhahn, T. Brox, D. Cremers, and H.-P. Seidel. A comparison of shape matching methods for contour based pose estimation. In R. Reulke, U. Eckhardt, B. Flach, U. Knauer, and K. Polthier, eds., *Combinatorial image analysis, 11th International Workshop, IWCIA 2006*, Berlin, Germany, 2006, *LNCS 4040*, pp. 263–276. Springer.
- [9] B. Rosenhahn, T. Brox, U. Kersting, A. Smith, J. Gurney, and R. Klette. A system for marker-less motion capture. *Künstliche Intelligenz*, 20(1):45–51, January 2006.
- [10] B. Rosenhahn, T. Brox, and J. Weickert. Three-dimensional shape knowledge for joint image segmentation and pose estimation. In W. Kropatsch, R. Sablatnig, and A. Hanbury, eds., *Pattern recognition, 27th DAGM Symposium*, Vienna, Austria, September 2005, *LNCS 3663*, pp. 109–116. Springer.
- [11] B. Rosenhahn, H. Ho, and R. Klette. Texture driven pose estimation. In M. Sarfraz, Y. Wand, and E. Banissi, eds., *Computer Graphics, Imaging and Visualization, New Trends (CGIV 05)*, Beijing, China, 2005, pp. 271–277. IEEE.
- [12] B. Rosenhahn, U. Kersting, K. Powell, R. Klette, G. Klette, and H.-P. Seidel. A system for articulated tracking incorporating a clothing model. *Machine Vision and Applications*, 18(1):25–40, 2007.
- [13] B. Rosenhahn, U. Kersting, K. Powell, and H.-P. Seidel. Cloth x-ray: MoCap of people wearing textiles. In K. Franke, K. R. Müller, B. Nickolay, and R. Schäfer, eds., *Pattern Recognition, 28th DAGM Symposium*, Berlin, Germany, September 2006, *LNCS 4174*, pp. 495–504. Springer.

- [14] B. Rosenhahn, U. Kersting, A. Smith, J. Gurney, T. Brox, and R. Klette. A system for marker-less human motion estimation. In W. Kropatsch, R. Sablatnig, and A. Hanbury, eds., *Pattern recognition, 27th DAGM Symposium*, Vienna, Austria, September 2005, *LNCS 3663*, pp. 230–237. Springer.
- [15] B. Rosenhahn, C. Perwass, and G. Sommer. Pose estimation of free-form contours. *International Journal of Computer Vision*, 62(3):267–289, May 2005.

15.9 Multiview Video Processing and Vision-based Computer Graphics

Coordinator: Christian Theobalt

This research group investigates problems that live on the boundary between the fields Computer Graphics and Computer Vision. One major line of research in Computer Vision aims at developing methods for acquiring dynamic scenes with video cameras and estimating model descriptions of the scenes from the recorded data. These model descriptions typically comprise of models of shape, models of motion or models of physical material properties. The main goal of Computer Graphics, on the other hand, has been to display such model descriptions realistically. In our work, we investigate the problems of acquisition, reconstruction and display of dynamic real-world scenes in conjunction and develop novel algorithmic concepts for each of these questions. In particular we develop new Vision and Graphics methods for dynamic shape and appearance reconstruction, motion estimation, animation of complex deformable models, real-time rendering and relighting, as well as fast image-and video processing. Ultimately, the combination of the developed techniques will enable us to generate realistic renderings of image- or video-captured dynamic scenes from arbitrary virtual camera views.

Numerous applications exist for such vision-based graphics algorithms. One prominent example is the emerging field of 3D Video and 3D TV in which the technological foundations for the next generation of visual media are developed. Here, the goal is to provide the viewer with the possibility to interactively change his viewpoint onto displayed video content, thereby enabling an immersive viewing experience. The work of our group on 3D Video is part of the European Union's Network of Excellence on 3DTV which plays a leading role in the advancement of this technology.

In the following, we will describe individual research projects in our group in more detail. Sect. 15.9.1 explains our new model-based framework for creating and relighting free-viewpoint videos of human actors which has been the first of its kind in the literature. In the context of this project, we also developed new approaches to capture dynamic scene geometry, as well as to create personalized human avatars from multi-view video. In Sect. 15.9.2 we present new methods to rapidly animate and realistically deform laser-scanned humans without the use of skeleton that facilitate model-based 3D Video creation. Our algorithms are based on fast Laplacian and Poisson mesh deformation and enable realistic mesh animation from a variety of input data formats, including raw optical marker trajectories and marker-free optical motion capture data. By further augmenting these methods and combining them with a marker-less tracking approach, we can even reconstruct the arbitrarily deforming dynamic geometry of moving subjects wearing complex apparel (e.g. a kimono)

(a) (b) (c)

Figure 15.24: Relightable Free-Viewpoint Video: (a) Side-by-side comparison between an input image and a color-coded reconstructed normal map; time-varying surface details such as wrinkles were reconstructed completely passively. The other two images show rendered free-viewpoint videos under captured real-world illumination (b) and a synthetic lighting setup as it is typical for virtual environments (c).

from raw multi-view video data, Sect. 15.9.3. Sect. 15.9.4 describes a new method to automatically learn kinematic structures of arbitrary moving subjects which greatly facilitates motion capture dynamic scene reconstruction. Finally, while the previous Sections were primarily concerned with the reconstruction of larger-scale scenes, Sect. 15.9.5 explains a new approach to realistic facial expression analogy that can be used to map facial motion between scanned 3D faces as well as faces in 2D video streams. We conclude by explaining our work on fast GPU-based image-and video processing in Sect. 15.9.6

15.9.1 A Model-based Approach to Reconstruct and Render Relightable Free-viewpoint Videos of Human Actors

Investigators: Christian Theobalt and Naveed Ahmed

By means of passive optical motion capture and dynamic -texturing 3D videos of real-people can be reconstructed from multi-view video footage [4]. The core of our method is a silhouette-based analysis-through synthesis method that, without the use of optical markings, shape adapts a template human body model to match the appearance of an actor and tracks its motion from the input video footage.

To import real-world characters into virtual environments, however, it is not enough to be able to reproduce the omni-directional appearance of an actor under the lighting conditions that prevailed at acquisition time. Rather, surface reflectance properties must be known in order to adapt the appearance to changing lighting environments. We therefore extended our original model-based free-viewpoint video approach by a concept called dynamic reflectometry which enables us to recover time-varying surface reflectance properties from eight input video streams captured under calibrated lighting [5, 8, 6, 7]. It enables us to render 3D videos in real-time from arbitrary viewpoints and under arbitrary illumination, Fig. 15.24(b),(c).

Our dynamic reflectance model comprises of a spatially-varying BRDF model for the surface, as well as a time-varying normal map capturing changes in surface geometry, Fig. 15.24(a). New methods to improve multi-view texture-to-image registration, as well as to detect and compensate cloth shifting prior to reflectance estimation further enhance reconstruction quality. A new technique for spatio-temporal reflectance sharing has also been developed in order to obtain faithful reflectance estimates despite only a sparse set of recording cameras [3].

We furthermore extended our original model-based free-viewpoint video framework in two other ways. First, we invented a spatio-temporal reconstruction technique to capture changes in surface geometry of the template model over time from the input footage. To serve this purpose, we apply a more detailed error measure during reconstruction that jointly takes into account silhouette-consistence, photo-consistency and mesh quality. By this means, pose dependent shape changes in geometry can be reproduced lending a further improved rendering quality [1].

Second, we proposed a new method to reconstruct personalized human avatars for interactive virtual environments from short multi-view video clips of a moving individual [2]. Our algorithm enables faithful reconstruction of body geometry, as well realistic estimation of a hole-free static surface texture by merging information from multiple video time steps. Again, spatio-temporal processing allows for high-quality results despite temporal in-visibility of certain surface areas.

References

- [1] E. de Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel. Reconstructing human shape and motion from multi-view video. In *2nd European Conference on Visual Media Production (CVMP)*, London, UK, December 2005, pp. 42–49. The IEE.

- [2] N. Ahmed, E. de Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel. Automatic generation of personalized human avatars from multi-view video. In *VRST '05: Proceedings of the ACM symposium on Virtual reality software and technology*, Monterey, USA, December 2005, pp. 257–260. ACM.

- [3] N. Ahmed, C. Theobalt, and H.-P. Seidel. Spatio-temporal reflectance sharing for relightable 3d video. In A. Gagalowicz and W. Philips, eds., *Third International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, MIRAGE 2007*, INRIA Rocquencourt, France, March 2007, *LNCS 4418*, pp. 47–58. Springer.

[4] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *ACM Transactions on Graphics*, 22(3):569–577, July 2003. (Proc. ACM SIGGRAPH '03).

- [5] M. Magnor, M. Pollefeys, W. Matusik, G. Cheung, and C. Theobalt. Video-based rendering. In *ACM SIGGRAPH 2005 Course Notes*, Los Angeles, USA, July 2005. ACM SIGGRAPH.

- [6] C. Theobalt, N. Ahmed, H. Lensch, M. Magnor, and H.-P. Seidel. Enhanced dynamic reflectometry for relightable free-viewpoint video. Research Report MPI-I-2006-4-006, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, July 2006.

- [7] C. Theobalt, N. Ahmed, H. 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(3), May 2007.

- [8] C. Theobalt, M. Magnor, and H.-P. Seidel. 3D image analysis and synthesis at MPI informatik. In D. W. Fellner, ed., *Vision, Video and Graphics 2005 (VVG'05)*, Edinburgh, UK, July 2005, pp. 85–91. Eurographics.

15.9.2 Skeleton-less Animation of Laser-Scanned Characters

Investigators: Edilson de Aguiar and Christian Theobalt

Photo-realistic computer-generated animations of humans are amongst the most important visual effects in motion pictures, virtual environments, and computer games. In order to realistically animate a character, most graphics artists make use of a well-established but often inflexible set of tools that is based on kinematic animation skeletons. However, in order to obtain lifelike character motion with animation skeletons, and in order to produce realistic surface deformations, a high amount of manual interaction is unavoidable.

We therefore developed simple, intuitive and fast approaches to animate high-quality triangle meshes of characters as they are acquired with our full-body laser scanner. Our methods are based on mesh deformation. They merely require the specification of a handful of correspondences between an input motion description and the vertices of a model to be animated. A first algorithmic variant is based on Poisson mesh deformation [2, 3]. While it enables the creation of lifelike character poses at interactive frame rates, a per-time step registration step is required. A second algorithmic variant is based on Laplacian mesh editing and a new method to infer and interpolate rotational constraints [1]. The latter algorithm also produces realistic results at high frame rates and does not require a registration step, Fig. 15.25. We recently began the development of a tetrahedral deformation approach that will process even better volume preservation characteristics and allow for fast approximation of non-linear deformation effects, see also Fig. 15.5. Since in all approaches only linear systems need to be solved, we easily achieve interactive frame rates.

Our methods only require a minimum of manual interaction, implicitly generates realistic body deformations, and can jointly handle many different types of input data. Motion transfer between subjects of completely different shape and proportions is also feasible. Our algorithms allow us to create lifelike character animations skeleton-based optical motion capture data and even directly from raw marker trajectories. By using a marker-less mocap method (Sect. 15.9.1), body poses recorded on video can be faithfully transfered to 3D models of arbitrary human subjects. The latter shows the potential of our method for high-quality free-viewpoint video production, since detailed scanned meshes can now be straightforwardly employed as template models.

References

- [1] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Rapid animation of laser-scanned humans. In *IEEE Virtual Reality 2007*, Charlotte, USA, 2007, pp. 223–226. IEEE.
- [2] E. de Aguiar, R. Zayer, C. Theobalt, M. Magnor, and H.-P. Seidel. A framework for natural animation of digitized models. Research Report MPI-I-2006-4-003, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, July 2006.
- [3] E. de Aguiar, R. Zayer, C. Theobalt, M. Magnor, and H.-P. Seidel. Video-driven animation of human body scans. In *IEEE 3DTV Conference*, Kos Island, Greece, 2007. IEEE.

Figure 15.25: Left: Subsequent frames showing the female scan authentically performing a soccer kick. Right: Motion parameters are extracted from raw video footage of human performances, and thereafter, the respective body poses are mapped to the scans of other human subjects.

15.9.3 Marker-less Deformable Mesh Tracking for Human Shape and Motion Capture

Investigators: Edilson de Aguiar and Christian Theobalt

One of the most important and most difficult tasks in 3D Video processing of human actors and other moving subjects, is the reconstruction of the complex time-varying surface representation and its motion from the input camera views. In order to deal with this challenging problem, we developed an approach that uses a high-quality scan of a subject as shape model, and captures its motion from multi-view video by means of a fast Laplacian deformation approach and an image-based 3D correspondence estimation method. [1].

Without using optical markers and not relying on kinematic skeletons, as traditional motion capture methods, our algorithm can handle humans wearing arbitrary clothing, including wide t-shirts, skirts and even kimonos, Fig. 15.26. The subject's overall motion, as well as non-rigid surface deformations are captured by the same unifying framework. Furthermore, our algorithm can straightforwardly be applied to any type of subject for which a scan exists, e.g. animals. Since it is based on a static template, it preserves the connectivity of the mesh over time, which is particularly important when it comes to our envisioned 3D Video applications and dynamic shape encoding.

Being easy to implement, our algorithm is able to make a laser scan of a subject move and deform in the same way as its real-world counterpart in video. We demonstrated the performance and accuracy of the method using synthetic and captured real-world video sequences. To our knowledge, this is the first system in the literature that can passively capture people with arbitrarily complex apparel, as well as other subjects exhibiting complex surface deformation.

References

- [1] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less deformable mesh tracking for human shape and motion capture. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, USA, June 2007. IEEE.

Figure 15.26: Left: Our method realistically captures the motion and the dynamic shape of a woman wearing a Japanese kimono from only eight video streams. Right: Side-by-side comparisons between an input video frame and the reconstructed pose of the laser scan. The poses of the persons and even the deformations of complex apparel, like the kimono, are faithfully reproduced.

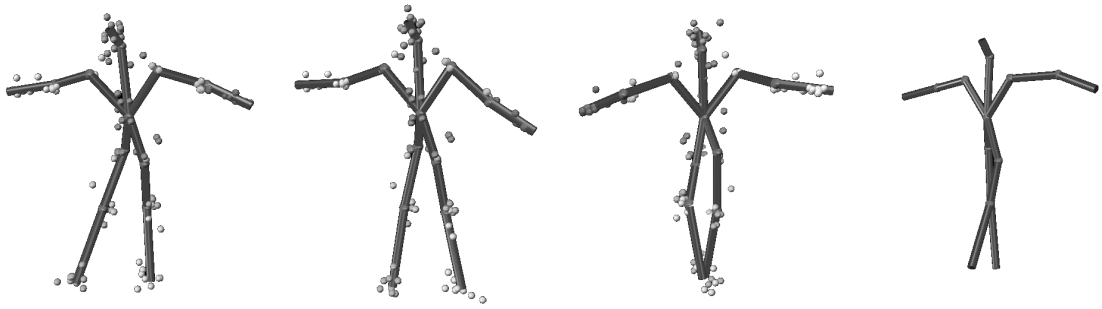


Figure 15.27: Skeleton reconstruction from a motion captured dancing sequence: The first three images show the motion capture markers and the learned skeleton in three different poses. The image on the right shows the skeleton with constant bone length at another time step. Joint positions and skeleton topology have been faithfully reconstructed.

15.9.4 Learning Kinematic Structures for Motion Analysis and Animation Processing

Investigators: Edilson de Aguiar and Christian Theobalt

In order to biomechanically analyze the motion of a person or in order to map real world performances onto virtual characters, captured marker-trajectories, e.g. 3D trajectories of optical beacons attached to the body, have to be transformed into the motion parameters of a kinematic skeleton model. Although commercial tools exist that assist professionals in performing this transformation, the estimation of kinematic skeletons and their motion parameters is still a labor-intensive, error prone and often inflexible process.

In order to automate this difficult task, we have developed a novel fully-automatic algorithm to estimate an articulated skeleton model of a moving subject and its motion parameters from body marker trajectories that were measured with an optical motion capture system [1] (Fig. 15.27). Our method does not require a priori information about the shape and proportions of the tracked subject, can be applied to arbitrary motion sequences, and renders dedicated initialization poses unnecessary. Our approach first identifies individual

rigid bodies by means of a variant of spectral clustering. Thereafter, it determines joint positions at each time step of motion through numerical optimization, reconstructs the skeleton topology, and finally enforces fixed bone length constraints. Through experiments, we validated the robustness and efficiency of our algorithm and showed that it outperforms related methods from the literature in terms of accuracy and speed. Our algorithm can also be applied to learn kinematic properties from mesh animations. A recent extension also enables kinematically plausible segmentation of moving meshes which renders useful during post-processing and compression of our captured deforming body models.

References

- [1] E. de Aguiar, C. Theobalt, and H.-P. Seidel. Automatic learning of articulated skeletons from 3D marker trajectories. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, and T. Malzbender, eds., *Advances in Visual Computing, Second International Symposium, ISVC 2006, Part I*, Lake Tahoe, NV, USA, November 2006, *LNCS 4291*, pp. 485–494. Springer.

15.9.5 A Generic Framework for 2D and 3D Facial Expression Analogy

Investigators: Zhao Dong and Christian Theobalt

Figure 15.28: VTC-based expression analogy: The first row and the second row represent the 3D and 2D expression analogy result, respectively .The left facial model/image pair in each row shows a source neutral image (l) and a source expressive image (r). The right facial model/image pair in each row shows a target neutral image (l) of another person that we made convincingly mimic (r) the expression of the source subject in the same row.

While 3D the previous projects were mainly concerned with 3D video processing of larger scenes involving moving humans, we also investigate techniques more specific to human faces.

Facial expression analogy provides computer animation professionals with a tool to map expressions of an arbitrary source face onto an arbitrary target face. This project aims at

providing a novel generic method for analogy-based facial animation that employs the same efficient framework to transfer facial expressions between arbitrary 3D face models, as well as between images of performer's faces [1] (Fig. 15.28). We propose a novel geometry encoding for triangle meshes, vertex-tent-coordinates(VTC), that enables us to formulate expression transfer in the 2D and the 3D case as a solution to a simple system of linear equations. Our experiments show that our method outperforms many previous analogy-based animation approaches in terms of achieved animation quality, computation time and generality.

Our algorithm is one further example of the fruitful combination of computer vision and computer graphics methods researched in our group. It has a variety of intriguing applications in animation and 2D/3D video processing. For instance, it enables us to map facial expression sequences from a video-sequence of one person to a photograph of another person.

References

[1] M. Song, Z. Dong, C. Theobalt, H. Wang, Z. Liu, and H.-P. Seidel. A generic framework for efficient 2D and 3D facial expression analogy. Technical report, Max-Planck-Institut für Informatik, 2007.

15.9.6 Real-time Image- and Video-processing on the GPU

Investigators: Gernot Ziegler and Christian Theobalt

Graphics processing units (GPUs) are data-parallel processors, capable of rapid analysis, transformation and compression of video streams. However, their hardware architecture is rather restricted, and makes the direct implementation of CPU software either impossible or highly inefficient. We take ideas from stream processing, computer graphics/vision and SIMD parallelization and combine them to yield GPU software prototypes which extend the realm of application from the mere post-processing to interactive computer graphics, real-time computer vision (e.g. robotics) and on-the-fly video compression.

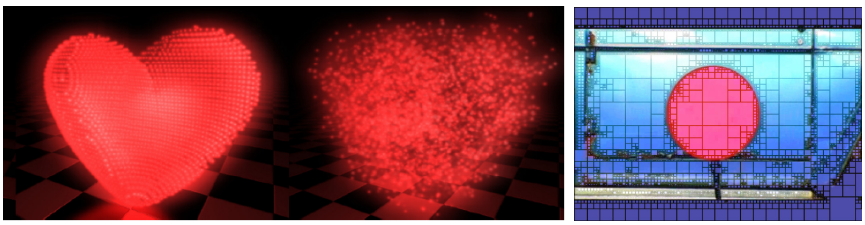


Figure 15.29: Left: An animated 3D model of a heart is converted into a point cloud, and explodes. Right: Video color-clustering through GPU quadtree analysis.

[3] enables dynamically sized lists of data on the GPU. It uses a pyramid of 2D data, similar to a mipmap, as an implicit indexing data structure to bypass the serial nature of list generation through pointer advancement. The method can be used to deliver visual effects for computer games, such as particle explosions of dynamic 3D models. Beyond this, the algorithm accelerates feature detection, pixel classification and binning, and enables sparse matrix handling, as detailed in [2].

[1] extended the building process for our 2D data pyramid with clustering abilities. This lead to clustering video or volume data into region quadtrees, fast enough to analyze PAL resolution video in real-time. Quadtree analysis becomes thus a light-weight preprocessing step for feature clustering in vision tasks, motion vector analysis, PDE calculations, or data compression.

References

- [1] G. Ziegler, R. Dimitrov, C. Theobalt, and H.-P. Seidel. Real-time quadtree analysis using histopyramids. In B. E. Rogowitz and T. N. Pappas, eds., *IS&T and SPIE Conference on Electronic Imaging*, San Jose, USA, January 2007, Proceedings of SPIE-IS&T Electronic Imaging. SPIE and IS&T.
- [2] G. Ziegler, A. Tevs, C. Theobalt, and H.-P. Seidel. GPU point list generation through histogram pyramids. Research Report MPI-I-2006-4-002, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, June 2006.
- [3] G. Ziegler, C. Theobalt, and H.-P. Seidel. On-the-fly point clouds through histogram pyramids. In L. Kobbelt, T. Kuhlen, T. Aach, and R. Westermann, eds., *11th International Fall Workshop on Vision, Modeling and Visualization 2006 (VMV2006)*, Aachen, Germany, 2006, pp. 137–144. Aka.

15.10 General Appearance Acquisition and Computational Photography

Coordinator: Hendrik Lensch

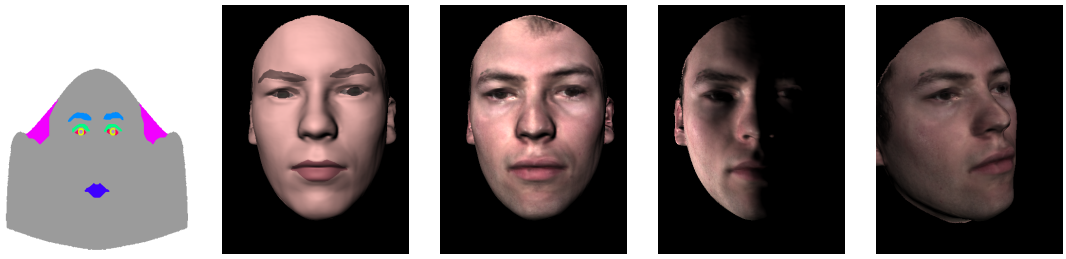
The goal of this group is to develop techniques and algorithms to capture realistic models of real-world objects and phenomena. The acquired models then allow for example to synthesize novel views of the digitized objects, to change the incident illumination virtually, or to extract specific attributes such as the object's fine-scale geometry or its spectral reflectance.

The acquisition setup makes use of optical systems augmented by computational procedures; by jointly designing the capturing apparatus, i.e., the optical layout of active and passive devices such as cameras, projectors, beam-splitters, etc., together with the capturing algorithm and appropriate post-processing. We use such combined systems to increase image quality, speed up acquisition time, to extract scene features, or to reconstruct complete 3D volumes from images.

15.10.1 Model-based Reflectance for Human Faces

Investigators: Martin Fuchs, Volker Blanz, and Hendrik P. A. Lensch

While reflectance fields represent a very general tool to model object appearance under varying illumination conditions, the involved data sets can grow to enormous proportions if the viewpoint and illumination conditions are to be chosen freely at rendering time. By restricting the modeled object class to human faces, we can greatly reduce the measurement effort, and require only about 20 input images in calibrated point light conditions as input [2]. Specifically, we employ the morphable model approach by Blanz and Vetter [1] to estimate



(a) (b) (c) (d) (e)

Figure 15.30: The face model allows to define a-priori regions (a) for which template BRDFs can be fitted (b) which are then refined to higher precision, enabling the rendering in novel light conditions and viewpoints (c), (d), (e).

a face geometry model and to establish correspondence between the input images. Then, we can fit a shift-variant BRDF model using a modification of the technique by Lensch et al. [3] (see Figure 15.30.)

References

[1] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063 – 1074, September 2003.

- [2] M. Fuchs, V. Blanz, H. P. A. Lensch, and H.-P. Seidel. Reflectance from images: A model-based approach for human faces. *IEEE Transactions on Visualization and Computer Graphics*, 11(3):296–305, May 2005.

[3] H. P. A. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics*, 22(2):234–257, April 2003.

15.10.2 Implicit Reflectance with Bayesian Relighting

Investigators: Martin Fuchs and Volker Blanz

The acquisition of precise reflectance models usually require carefully controlled measurement conditions. We can relax these requirements, and recover the appearance of a complex scene in novel illumination, by observing a reference object in this illumination [1].

We record a set of input images both of the scene (e.g. a human face) and a reference object (such as a snooker ball) in low-frequency, distant, but uncalibrated illumination. For an image of the reference object in novel illumination, we can then estimate a linear combination of the input reference images. Applying the linearity of light transport, the same linear combination for the input images of the scene are used to obtain a rendering in novel light. The estimation of the linear combination is sensitive to image noise; we therefore employ a Bayesian approach with a Gaussian noise model, and thus regularize the solution efficiently (see Figure 15.31).

$\eta = 0$ $\eta = 0.001$ $\eta = 0.01$ $\eta = 0.1$ $\eta = 1$ $\eta = 100$

Figure 15.31: Renderings of a face in novel light, with a close-up on the eye to the left. The regularization parameter η allows a continuous control of the trade-off between accurate reconstruction (low values) and increased noise (high values).

References

- [1] M. Fuchs, V. Blanz, and H.-P. Seidel. Bayesian relighting. In O. Deussen, A. Keller, K. Bala, P. Dutré, D. W. Fellner, and S. N. Spencer, eds., *Rendering Techniques 2005: Eurographics Symposium on Rendering*, Konstanz, Germany, July 2005, Rendering Techniques, pp. 157–164. Eurographics.

15.10.3 Near-Field Reflectance Fields

Investigator: Hendrik P. A. Lensch

Figure 15.32: Near-field reflectance field. From left to right: scene rendered from the camera's viewpoint, from the projector's view point, relit with a spatially varying pattern, simulated area light source after interpolating the reflectance samples.

Most current reflectance fields approaches assume distant light sources. In this project we have developed novel, efficient techniques for capturing reflectance fields which can be correctly relit even with spatially varying light patterns and near-by light sources. Our approach for the first time allows for accurate measurements of the ray-to-ray light transport in an arbitrarily complex scene. In our dual photography approach [3] we compute efficient parallelized and adaptive illumination patterns for capturing the pixel-to-pixel light transport between a camera and a projector. Our algorithm captures the scene about 1000 times faster than any previous approach. Based on the captured reflectance field and exploring Helmholtz reciprocity it is possible to swap the role of the camera and the projector virtually. The scene can be rendered realistically from the point of view of the projector while being illuminated with arbitrary patterns from the camera. Although no camera was present all

possible direct and global light transport paths are reproduced correctly including specular reflections, subsurface scattering, interreflections and caustics (see Figure 15.32). With our symmetric photography technique [2] we have further augmented the approach to capture the full 8D reflectance field from multiple cameras and multiple projectors of scenes with strong interreflections. Using $\mathcal{H}$-matrices we exploit the data-sparseness in dense transport matrices and achieve better SNR, faster acquisition and more compact compression compared to the dual photography approach.

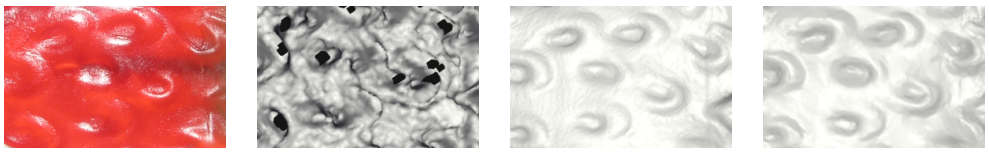
Finally, since the angular light resolution, i.e. the number of captured projector positions, in the symmetric photography is still relatively limited we have developed an interpolation scheme [1] that can produce realistic images even for in-between projector locations. It is now possible to faithfully reproduce the appearance of the scene for smoothly moving or extended light sources, generating both shadows and highlights correctly.

References

- [1] B. Chen and H. P. A. Lensch. Light source interpolation for sparsely sampled reflectance fields. In G. Greiner, J. Hornegger, H. Niemann, and M. Stamminger, eds., *Vision, Modeling, and Visualization 2005 (VMV'05)*, Erlangen, Germany, November 2005, pp. 461–469. Aka.
- [2] G. G. Garg, E.-V. Talvala, M. Levoy, and H. P. A. Lensch. Symmetric photography: Exploiting data-sparseness in reflectance fields. In T. Anenine-Möller and W. Heidrich, eds., *Rendering Techniques 2006: Eurographics Symposium on Rendering*, Nicosia, Cyprus, June 2006, pp. 251–262. Eurographics Association.
- [3] P. Sen, B. Chen, G. G. Garg, S. Marschner, M. Horowitz, M. Levoy, and H. P. A. Lensch. Dual photography. *ACM Trans. on Graphics (Proc. SIGGRAPH 2005)*, 24(3):745–755, August 2005.

15.10.4 Mesostructure from Specularity

Investigators: Tongbo Chen and Michael Goesele



(a) (b) (c) (d)

Figure 15.33: Mesostructure reconstruction of a piece of jelly candy (ca. 18mm×28mm). (a) Input image. (b) Shape by laser scanner. (c) Shape by laser scanner after covering the object with Lambertian powder, which is close to ground truth, but missing some details. (d) The shape by our method reveals more fine details.

This project aims to provide a simple and robust method for surface mesostructure acquisition. Our method builds on the observation that specular reflection is a reliable visual cue for surface mesostructure perception. In contrast to most photometric stereo methods, which take specularities as outliers and discard them, we propose a progressive acquisition system that captures a dense specularity field as the only information for mesostructure

reconstruction. Our method can efficiently recover surfaces with fine-scale geometric details from complex real-world objects with a wide variety of reflection properties, including translucent, low albedo, and highly specular objects. We show results for a variety of objects including skin, apricot, orange, jelly candy, black leather and dark chocolate.

References

- [1] T. Chen, M. Goesele, and H.-P. Seidel. Mesostructure from specularity. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, New York, NY, USA, 2006, vol. 2, pp. 1825–1832. IEEE.

15.10.5 Acquiring Multiperspective Street Panoramas

Investigator: Hendrik P. A. Lensch

Figure 15.34: Multiperspective street panorama: push-broom image (top) with severe distortions, optimized perspectives (middle) and a larger section of the entire street (bottom).

City maps typically only provide information about the street names and the layout of the streets. The appearance of the street at some location is typically not communicated. In this joint project with Google we have therefore investigated how an entire street can be efficiently digitized and represented by a single multiperspective image [1]. The goal is to immerse the user partially into the real scene allowing for new means of navigation through the city, e.g. visually locating a specific shop or restaurant if the exact address is unknown. Driving by a street, using a time-of-flight scanner, a rough approximation of the 3D geometry is captured. At the same time a HDR video stream is recorded with a high speed camera. After registering the image content with the 3D geometry one could render a textured 3D model of the street. However, a 3D model is difficult to transmit and to browse at an arbitrary client PC. Therefore, we generate a single summarizing multiperspective picture of the entire street. Multiple perspectives are necessary because of the extreme width of the scene. While the perspective in the vertical direction is given by the capturing camera we can create different horizontal perspectives by combining the information of consecutive frames. If one selects an orthographic projection in x a so-called push-broom image is created. For surface points that are not at the focus plane the aspect ratio is distorted leading to severe artifacts as seen in the Figure 15.34. Our algorithm creates a series of smooth blends between different perspectives for the entire length, trying to minimize these distortions. For scene parts with high depth variation such as intersections we try to approximate the original perspective in x while for scene parts at the focus plane, i.e. the store fronts, we are free to

pick any perspective to provide a smooth blend. Our algorithm [1] performs the optimization hierarchically and can efficiently optimize extremely long panoramas such as the street shown in Figure 15.34 which is 2km in length.

References

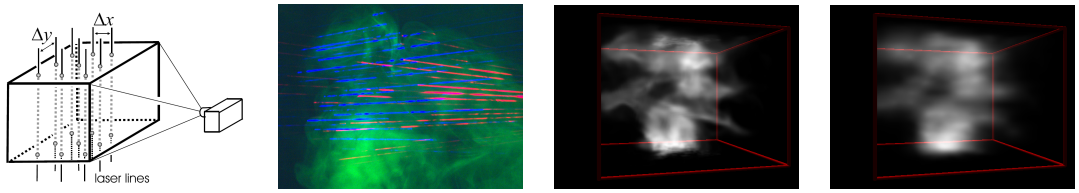
- [1] A. Román and H. P. A. Lensch. Automatic multiperspective images. In T. Akenine-Möller and W. Heidrich, eds., *Rendering Techniques 2006: Eurographics Symposium on Rendering*, Nicosia, Cyprus, June 2006, pp. 161–171. Eurographics Association.

15.10.6 Volume Density Acquisition

Investigators: Christian Fuchs, Tongbo Chen, Michael Goesele, and Holger Theisel

We acquire spatially varying densities in a volume of participating media (such as smoke) instantaneously with a single image. The volume is illuminated with one or more sets of laser lines and the scattered intensity is recorded using a conventional camera. This approach sacrifices continuous sampling in the spatial domain in favor of continuous sampling in the time domain. Density information is extracted along every laser line. The projection of each line onto the image plane is determined by geometric calibration of the measurement setup. Multiple scattering which would otherwise destroy the measurements, is removed.

Different techniques can be employed in order to reconstruct a complete volume from the densities that were measured along the laser lines. Simple approximation using a cosine-kernel for example yields smooth results but obviously averages out high-frequency detail. More advanced methods like the push-pull algorithm by [1] generally produce better results by preserving the high frequencies but sometimes introduce visible artifacts due to the non-uniform sampling.



(a) (b) (c) (d)

Figure 15.35: (a) shows our measurement setup. The lasers and the camera are arranged such that all lasers in one set are projected without occlusions onto the image plane. The resulting picture is shown in (b). We performed an analysis of the reconstruction performance using simulated ground truth data (c). Reconstruction using 2 sets of 10×10 lasers results in (d).

References

[1] I. Drori, D. Cohen-Or, and H. Yeshurun. Fragment-Based Image Completion. *ACM Transactions on Graphics*, 22(3):303–312, 2003.

- [2] C. Fuchs, T. Chen, M. Goesele, H. Theisel, and H.-P. Seidel. Volumetric density capture from a single image. In T. Möller, R. Machiraju, M.-S. Chen, and T. Ertl, eds., *Volume Graphics 2006, Eurographics / IEEE VGTC Workshop Proceedings*, Boston, USA, 2006, pp. 17–22. Eurographics.

15.10.7 HDR Demosaicing and Multispectral Imaging

Investigators: Boris Ajdin and Hendrik P. A. Lensch

Figure 15.36: Scene captured at 5 different wavelengths.

Most current digital cameras capture only three primary colors using a color filter array (cfa) in front of the imaging sensor in such a way that each original pixel captures only one color channel. The missing two channels need to be interpolated from the near-by pixels in a process typically called demosaicing. While a large number of demosaicing algorithms have been developed for low-dynamic range images they are inaccurate for HDR content where the contrast for neighboring pixels can be significantly higher. In this group we therefore investigate novel demosaicing algorithms with better interpolation properties for HDR images.

We have furthermore assembled a system for capturing multispectral HDR images by combining a tunable color filter with a cfa camera. The tunable filter extracts wavelength intervals of approximately $7nm$, and we capture 31 images with wavelength regularly spaced over the visual spectrum. For each selected wavelength interval the different colors in the color filter array now act as neutral density filters which can be used to extend the dynamic range of a single exposure capture. We plan to use this multispectral camera to investigate reflection properties at different wavelength much more precisely than our previous RGB measurements.

15.11 Advanced Global Illumination and Realtime Realistic Image Synthesis

Coordinators: Johannes Günther and Karol Myszkowski

Ray tracing and global illumination techniques significantly improve the realism of synthetic images. Both the hardware and algorithms at hand are now mature enough to make realistic rendering of animated environments feasible even at interactive speeds.

In Section 15.11.1 we describe our work which improves the performance as well as the quality of general global illumination algorithms based on photon density estimation. Because ray shooting is very often the core technique for global illumination methods we need to enhance its flexibility and performance. The results of our research in this direction is

summarized in Section 15.11.2. We developed solutions to efficiently support more primitives for ray tracing and, moreover, lift the limitation to static scenes by proposing several novel algorithms to ray trace changing scenes interactively. Section 15.11.3 deals with exploiting the temporal coherence which is present when rendering animations, shortening the overall rendering time by simultaneously improving quality. Finally, we describe the physically-based simulation of the twilight phenomena in Section 15.11.4.

15.11.1 Global Illumination with Photon Density Estimation

Investigators: Vlastimil Havran and Robert Herzog

Photon mapping is a widely used global illumination algorithm because it can efficiently compute all kind of lighting effects in a unified framework. In photon mapping, incident radiant flux (photons) is cached in an initial propagation pass, which is used in the following rendering pass to estimate the outgoing radiance at specific points on the scene surfaces. Since the direct radiance estimate from the local photon density is rather poor, stochastic Monte Carlo sampling of the local BRDF function is initiated to integrate the incident radiance from the surrounding surfaces. This process is called *final gathering*. Final gathering requires to shoot hundreds to thousands of sample rays (final gather rays) per pixel in order to obtain results with low variance. For each final gather ray the radiance is estimated from the local photon density computed via k-nearest neighbor density estimation. Final gathering and searching for the k-nearest neighbor photons is the computationally most expensive operation during rendering.

In [2], we presented a new algorithm for computing high quality indirect illumination based on photon density estimation similarly to photon mapping. We accelerate the search for final gathering by reorganizing the computation in the reverse order. We make use of a dual-tree approach that spatially organizes the positions of photons in one tree and the nearest hit points of the final gather rays in another tree, referred to as *reverse photons*. Instead of searching for the k-nearest photons in the neighborhood of a reverse photon, we distribute the photon's energy weighted by a bivariate kernel to all contributing reverse photons and their corresponding pixels. The performance improvement takes advantage of the logarithmic complexity of searching in trees. We further demonstrate how the algorithm is combined with other acceleration techniques such as *photon density control* (Figures 15.37a – c) and our novel *caching of final gather rays* for coherent final gather ray shooting. The whole algorithm is cache-oblivious as it creates coherent access patterns to the main memory, which further improves the performance and also allows us to use efficient external data structures to alleviate the increased memory requirements.

Despite its generality, the photon map has one major drawback: the radiant energy carried by the photons can only be measured in proximity of the photon impacts and the photon density is computed over a finite neighborhood. This generates biased results in particular near the boundaries of objects. In [1] we presented a novel data structure called *ray map* that partially solves the limitations of photon mapping. The ray map extends the concept of photon maps: it stores not only photon impacts but the whole photon paths represented by a sequence of rays. Ray maps are used for k-nearest neighbor density estimation similarly to photon maps. However, ray maps represent 4D information while photon maps store only 3D information of the light transport. This avoids visible bias in the photon density estimate near

(a) (b) (c) (d) (e)

Figure 15.37: Computing global illumination images using photon density estimation. (a) photon distribution without density control, (b) with density control, (c) the resulting high-quality image for indirect illumination in the Apartment scene© INRIA 2005 rendered with reverse photon mapping. The results obtained using (a) or (b) as input are visually indistinguishable, while the number of photons in (b) is reduced by a factor of two. The colored points in image (a) and (b) represent the photon records with color-coded splatting bandwidth proportional the local photon density. Images (d) and (e) show a direct visualization of the photon map and our proposed ray map data structure respectively both rendered with k-NN photon density estimation. Note the boundary bias in image (d).

surface boundaries and also reduces bias on surfaces with complex geometric topology (refer to Figures 15.37d and e). Since the ray map indexes entire photon paths rather than photon hits, it requires more memory resources. Therefore, we propose a particular representation of ray maps using a lazily constructed spatial subdivision based on kd-trees.

References

- [1] V. Havran, J. Bittner, R. Herzog, and H.-P. Seidel. Ray maps for global illumination. In O. Deussen, A. Keller, K. Bala, P. Dutré, D. W. Fellner, and S. N. Spencer, eds., *Rendering Techniques 2005: Eurographics Symposium on Rendering*, Konstanz, Germany, June 2005, pp. 43–54,311. Eurographics.

- [2] V. Havran, R. Herzog, and H.-P. Seidel. Fast final gathering via reverse photon mapping. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, August 2005, *Computer Graphics Forum*, vol. 24, pp. 323–333. Blackwell.

15.11.2 Interactive Ray Tracing of Dynamic Scenes

Investigators: Johannes Günther, Ingo Wald, Vlastimil Havran, Robert Herzog, Alexander Efremov, Tongbo Chen, and Michael Goesele

Many global illumination algorithms – including photon mapping – use ray tracing as their core technique to determine visibility. Thus making the ray tracing function faster and more flexible will directly improve these global illumination methods as well.

To increase the utility of ray tracing we developed techniques to efficiently support other primitives. Besides spheres or triangle meshes we are able to also ray trace iso-surfaces [9],

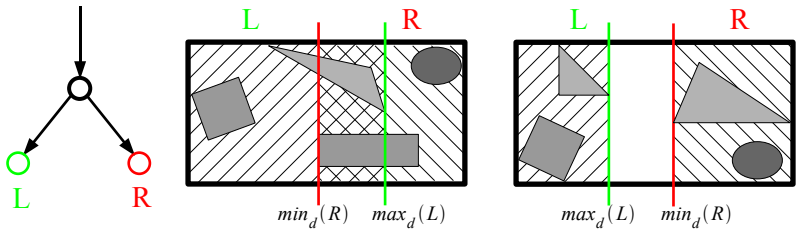


Figure 15.38: An organization of splitting planes inside a SKD-node (left). Two splitting planes in the node define the spatial extent of two children: (middle) the child nodes overlap, (right) the child nodes are disjoint.

NURBS [1], or surfaces represented by points [10] at interactive frame rates.

Application for interactive ray tracing are in games [2] or in the (automotive) industry [8]. In the latter case an important issue is the reflectance function (BRDF) acquisition and rendering validation against measurement data [6, 3].

Though ray tracing has recently become interactive, its high precomputation time for building spatial indices (which is in $O(N \log N)$) usually limits its applications to walk-throughs of static scenes. This is a major limitation, as most applications demand support for dynamically animated models.

There are basically three strategies for interactive ray tracing of dynamic scenes. First, we can optimize other acceleration structures for ray tracing that can be quickly built or updated, in contrast to the usually used kd-trees. Second, we can avoid the costly re-build of kd-trees. And third, one can speed-up and optimize the *construction* of kd-trees. We followed all these approaches and detail our solutions in the next paragraphs.

We start with showing how spatial kd-trees (SKD-trees) can be used for ray tracing instead of standard kd-trees. In contrast to classic kd-trees, SKD-trees allow for overlapping of spatial regions by using two axis-aligned splitting planes per node (refer to Figure 15.38). Secondly, we show that it is convenient to combine the spatial kd-trees with bounding volumes in a sparse way, resulting in a hybrid data structure with $O(N \log N)$ preprocessing time and $O(N)$ storage. We also lift the assumption on $O(N \log N)$ time complexity for preprocessing to $O(N \log \log N)$ by discretization in the space of splitting planes. Radix sort (bucket/distribution sort etc.) which relies on the limited precision of input data, achieves $O(N)$ time complexity for sorting instead of $O(N \log N)$ based on comparisons. Similarly, we construct an efficient spatial hierarchy for ray tracing using a cost model based on the surface area heuristic (SAH) in a discrete setting. The construction assumes that the representation of axis-aligned bounding boxes, is restricted to b bits. Furthermore our method assumes that the objects do not vary much in size and that the distribution of objects is not highly skewed. The experimental evaluation of our method on six scenes of various complexity, revealed an average total speedup, including spatial hierarchy construction and ray tracing, of factor 4 compared with standard kd-trees. We achieve about 3 to 5 frames per second for rendering the BART animation test scenes with ray tracing using our proposed data structure on standard PC hardware.

Next, we present a new approach to ray trace a special but important class of dynamic scenes, namely models whose connectivity does not change over time and for which all possible poses are known in advance [5]. We support these kinds of models by introducing two new concepts: motion decomposition, and fuzzy kd-trees. We analyze the animation and break the model down into submeshes with similar motion (see also Figure 15.39 left). For each of these submeshes and for every time step, we calculate a best affine transformation through a least square approach. Any residual motion is then captured in a *single* "fuzzy kd-tree" for the entire animation. Together, these techniques allow for ray tracing animations *without* rebuilding the spatial index structures for the submeshes, resulting in interactive frame rates of 5 to 15 fps even on a single CPU.

Although we can now ray trace animations interactively we are still limited to animations where all poses are known in advance. In [4], we ease this limitation considerably for the case of skinned models as typically used in computer games. We assume that the characters are built from meshes with an underlying skeleton structure, where the set of joint angles defines the character's pose and determines the skinning parameters. Based on a sampling of the possible pose space we build a static fuzzy kd-tree for each skeleton segment in a fast preprocessing step. These fuzzy kd-trees are then organized into a top-level kd-tree. Together with the skeleton's affine transformations this multi-level kd-tree allows fast and efficient scene traversal at runtime, while arbitrary combinations of animation sequences can be applied interactively to the joint angles. A ray traced example scene including shadows is shown in Figure 15.39 right. We achieve interactive frame rates including shadows at 1024×1024 resolution even on a single processor core.

The former strategies for interactive ray tracing of dynamic scenes try to avoid the costly re-build of kd-trees as much as possible. This, however, limits the type of motion present in the animation: random motion of triangles, for example, cannot be handled. To address this problem we also developed methods to speed-up and optimize the *construction* of (SAH based) kd-trees [7]. We propose modifications that significantly increase the coherence of memory accesses during construction of the kd-tree. Additionally we provide theoretical and practical results regarding *conservatively* sub-sampling of the SAH cost function. Using this method we achieve order of magnitude faster build times than previous published results, making it feasible to build optimized kd-trees for significantly larger dynamic scenes from scratch every frame.

References

- [1] A. Efremov, V. Havran, and H.-P. Seidel. Robust and numerically stable Bézier clipping method for ray tracing NURBS surfaces. In *SCCG '05: Proceedings of the 21st Spring Conference on Computer Graphics*, Budmerice, Slovakia, May 2005, pp. 127–135. ACM.
- [2] H. Friedrich, J. Günther, A. Dietrich, M. Scherbaum, H.-P. Seidel, and P. Slusallek. Exploring the use of ray tracing for future games. In *ACM SIGGRAPH Video Game Symposium sandbox '06: Proceedings of the 2006 ACM SIGGRAPH symposium on Videogames*, Boston, MA, USA, July 2006, pp. 41–50. ACM.
- [3] J. Günther, T. Chen, M. Goesele, I. Wald, and H.-P. Seidel. Efficient acquisition and realistic rendering of car paint. In G. Greiner, J. Hornegger, H. Niemann, and M. Stamminger, eds., *Vision, Modeling, and Visualization 2005 (VMV'05)*, Erlangen, Germany, November 2005, pp. 487–494. Aka.

Figure 15.39: Left: The CHICKEN animation ray traced in realtime. The color encodes separated clusters generated by our motion decomposition algorithm. Right: The CALLY example scene demonstrating interactively skinned meshes with ray traced shadows at 1024×1024 pixels on a single CPU.

- [4] J. Günther, H. Friedrich, H.-P. Seidel, and P. Slusallek. Interactive ray tracing of skinned animations. *The Visual Computer*, 22(9-11):785–792, September 2006.
- [5] J. Günther, H. Friedrich, I. Wald, H.-P. Seidel, and P. Slusallek. Ray tracing animated scenes using motion decomposition. *Computer Graphics Forum*, 25(3):517–525, September 2006.
- [6] V. Havran, A. Neummann, G. Zotti, W. Purgathofer, and H.-P. Seidel. On cross-validation and resampling of brdf data measurements. In *SCCG '05: Proceedings of the 21st Spring Conference on Computer Graphics*, Budmerice, Slovakia, May 2005, pp. 161–168. ACM.
- [7] S. Popov, J. Günther, H.-P. Seidel, and P. Slusallek. Experiences with streaming construction of SAH kd-trees. In I. Wald and S. G. Parker, eds., *Proceedings of the 2006 IEEE Symposium on Interactive Ray Tracing*, Salt Lake City, USA, September 2006, pp. 89–94. IEEE. Best Paper Award.
- [8] I. Wald, A. Dietrich, C. Benthin, A. Efremov, T. Dahmen, J. Günther, V. Havran, H.-P. Seidel, and P. Slusallek. A ray tracing based framework for high-quality virtual reality in industrial design applications. In I. Wald and S. G. Parker, eds., *Proceedings of the 2006 IEEE Symposium on Interactive Ray Tracing*, Salt Lake City, USA, September 2006, pp. 177–185. IEEE.
- [9] I. Wald, H. Friedrich, G. Marmitt, P. Slusallek, and H.-P. Seidel. Faster isosurface ray tracing using implicit kd-trees. *IEEE Transactions on Visualization and Computer Graphics*, 11(5):562–572, September 2005.
- [10] I. Wald and H.-P. Seidel. Interactive ray tracing of point-based models. In M. Pauly and M. Zwicker, eds., *Symposium on Point-Based Graphics*, Stony Brook, USA, June 2005, pp. 9–16. Eurographics Association.

15.11.3 Exploiting Temporal Coherence in Animation Rendering

Investigators: Vlastimil Havran, Grzegorz Krawczyk, and Karol Myszkowski

Producing high quality animations featuring rich object appearance and compelling lighting effects is very time consuming using traditional frame-by-frame rendering systems. In

this section we present our global illumination and rendering solutions that exploit temporal coherence in lighting distribution for subsequent frames to improve the computation performance and overall animation quality [3, 2, 1].

In animation rendering the irradiance cache computation is a real computation bottleneck [4] that prevents a more widespread use of global illumination effects in a standard movie production pipeline. The main goal of our research [3] is the improvement of the irradiance cache performance through exploiting temporal coherence between cache locations and their irradiance samples for the subsequent animation frames. This naturally leads to a significant reduction of popping artifacts as well. We introduce a lazily build data structure, which monitors local changes of illumination across the scene at the so-called anchor points, whose density adapts to changes in the lighting distribution. Since anchor points are directly linked to hemispherical samples (strata) for each irradiance cache, lighting updated at those points for each frame is immediately propagated to the corresponding strata (many strata belonging to different caches can be linked to the same anchor). A visibility map computed using graphics hardware is used to detect occlusions/disocclusions between anchors and a given cache location, which are caused by dynamic objects in the scene. All affected strata are identified in the visibility map and ray tracing is used to assign another nearest anchor point or possibly to create a new one. Thus, in our approach costly ray tracing is only used to update visibility changes caused by dynamic objects and it is not repeated for all strata and for each frame as in traditional approaches.

Realistic rendering of smoke, dust, and fog effects at interactive speeds is required in many applications, but it is expensive in particular for dynamic participating media. Again, we exploit temporal coherence in the global illumination computation using selective photon tracing, which enables us to update mostly those photon paths that are affected by changes in the media [2]. We enhance the selective photon tracing technique by eliminating the need of shooting corrective photons, which enables to achieve interactive framerates.

Since incorporating captured real world lighting into rendering pipeline results in extreme realistic rendering, we propose an interactive system for fully dynamic scene lighting using captured high dynamic range (HDR) video environment maps [1]. The key component of our system is an algorithm for efficient decomposition of HDR video environment map captured over hemisphere into a set of representative directional light sources, which can be used for the direct lighting computation with shadows using graphics hardware (refer to Figure 15.40). The resulting lights exhibit good temporal coherence and their number can be adaptively changed to keep a constant framerate while good spatial distribution (stratification) properties are maintained. We can handle a large number of light sources with shadows using a novel technique which reduces the cost of BRDF-based shading and visibility computations. We demonstrate the use of our system in a mixed reality application in which real and synthetic objects are illuminated by consistent lighting at interactive framerates.

References

- [1] V. Havran, M. Smyk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Interactive system for dynamic scene lighting using captured video environment maps. In O. Deussen, A. Keller, K. Bala, P. Dutré, D. W. Fellner, and S. N. Spencer, eds., *Rendering Techniques 2005: Eurographics Symposium on Rendering*, Konstanz, Germany, June 2005, pp. 31–42,311. Eurographics.

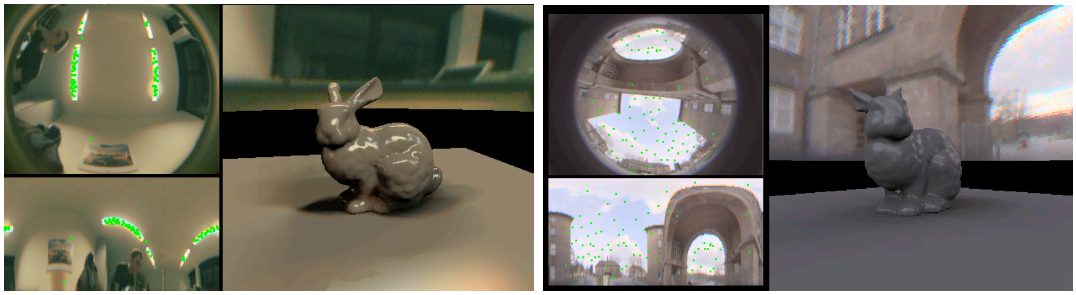


Figure 15.40: Left: the model with strong specular reflectance properties composed of 16,200 triangles rendered with 72 shadow maps at 5.3 Hz. On the left top is an environment map captured in real-time using an HDR camera with fisheye lens. The light sources are marked by green points. On the left bottom the same environment map is shown in polar projection. Right: the same model illuminated by outdoor lighting.

- [2] J.-R. Jiménez, K. Myszkowski, and X. Pueyo. Interactive global illumination in dynamic participating media using selective photon tracing. In *SCCG '05: Proceedings of the 21st Spring Conference on Computer Graphics*, Budmerice, Slovakia, 2005, pp. 211–218. ACM.
- [3] M. Smyk, S. Kinuwaki, R. Durikovic, and K. Myszkowski. Temporally coherent irradiance caching for high quality animation rendering. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, 2005, *Computer Graphics Forum*, vol. 24, pp. 401–412. Blackwell.

[4] E. Tabellion and A. Lamorlette. An Approximate Global Illumination System for Computer-Generated Films. *ACM Transactions on Graphics*, 23(3):469–476, 2004.

15.11.4 Physically-based Simulation of Natural Optical Phenomena

Investigators: Jörg Haber and Marcus Magnor

Rendering outdoor scenes almost always includes vistas of the sky. So far, either actual photographs or parametric skylight models had to be used for that purpose. But especially for twilight daytimes with their attractive, quickly changing and diverse lighting conditions, no suitable sky modeling tool has been available. In particular, for animations of a sunrise/sunset, standard photographs/video cannot reproduce the dynamic range of the sky while HDR images might be problematic to acquire in the short time of constant sky colors, and parametric models cannot always deliver the subtle nuances of twilight phenomena.

We present a physically based approach to compute the colors of the sky during the twilight period before sunrise and after sunset [1]. The simulation is based on the theory of light scattering by small particles. A realistic atmosphere model is assumed, consisting of air molecules, aerosols, and water. Air density, aerosols, and relative humidity vary with altitude. In addition, the aerosol component varies in composition and particle size distribution. This allows us to realistically simulate twilight phenomena for a wide range of different climate conditions. Besides considering multiple Rayleigh and Mie scattering, we take into account wavelength-dependent refraction of direct sunlight as well as the shadow of the Earth. Incorporating several optimizations into the radiative transfer simulation, a photo-realistic

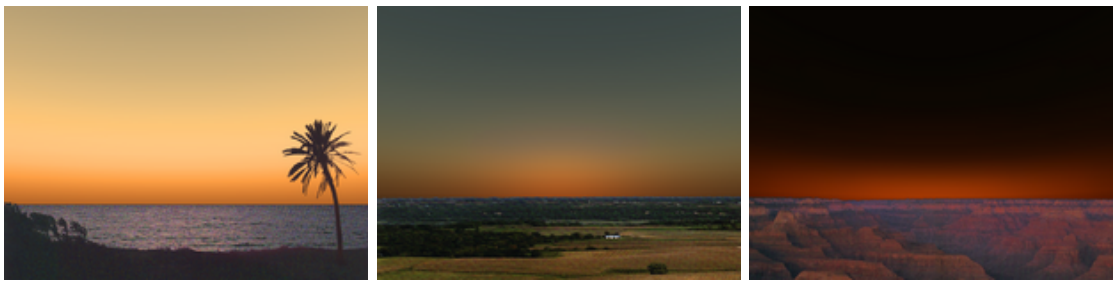


Figure 15.41: Various twilight phenomena simulated with our system for different times and climates. Left to right: horizontal stripes appear a few minutes after sunset (sun elevation $-0.5°$ below the horizon, maritime climate); purple light is at its strongest about 20–30 minutes after sunset (sun elevation $-30°$, continental climate); afterglow shows up about 40–45 minutes after sunset (sun elevation $-6°$, continental climate). Nomenclature of twilight phenomena according to Minnaert.

hemispherical twilight sky is computed in less than two hours on a conventional PC (refer to Figure 15.41). The resulting radiometric data is useful, for instance, for high-dynamic range environment mapping, outdoor global illumination calculations, mesopic vision research and optical aerosol load probing.

References

- [1] J. Haber, M. Magnor, and H.-P. Seidel. Physically based simulation of twilight phenomena. *Transactions on Graphics*, 24(4):1353–1373, October 2005.

15.12 High Dynamic Range Imaging and Perception Issues in Graphics

Coordinators: Rafał Mantiuk and Karol Myszkowski

The goal of this group is to develop algorithms for capturing, processing, storing and display of images and video that preserve the colors and the luminance range 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.

Vast majority of digital images and video material stored today can capture only a fraction of visual information visible to the human eye and does not offer sufficient quality to reproduce them on the future generation of display devices. The limiting factor is small color gamut and even more limited dynamic range (contrast) that cameras can capture and that majority of image and video formats can store. The gap between the range of luminance and color gamut that can be perceived by the human eye and that can be encoded with the existing formats is illustrated in Figure 15.42.

High dynamic range imaging (HDRI) addresses the shortcoming of the traditional imaging by capturing, processing and storing visual information with much higher precision. In this

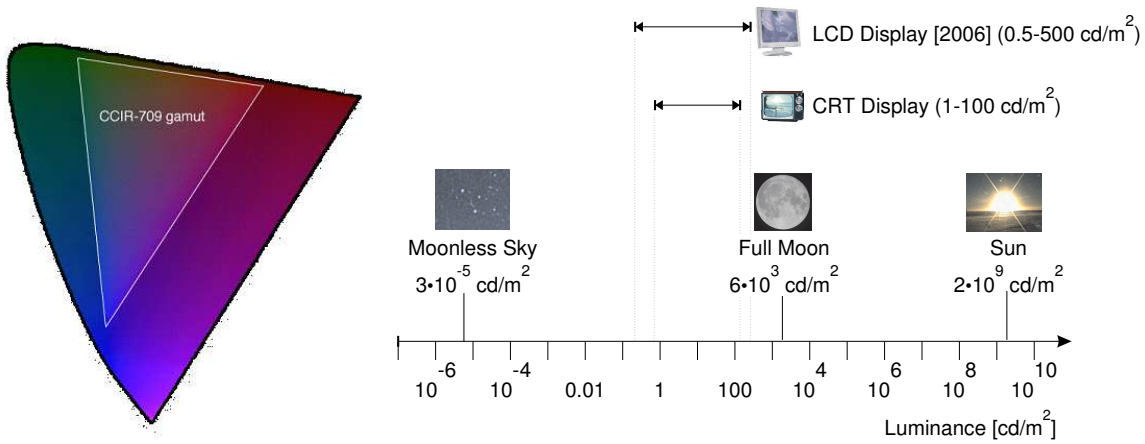


Figure 15.42: Left: color gamut frequently used in traditional imaging (CCIR-705), compared to the full visible color gamut. Right: real-world luminance values compared with the range of luminance that can be displayed on CRT and LDR monitors.

group we take advantage of such high dynamic representation to develop algorithms for efficient storage and display of high fidelity digital images and video.

15.12.1 High Dynamic Range Image and Video Compression

Investigators: Rafał Mantiuk, Alexander Efremov, and Karol Myszkowski

Since popular image and video compression formats, such as MPEG (ISO/IEC 14496-2, -10) or JPEG, do not offer sufficient precision to encode high dynamic range content, we have developed several encoding algorithms for high dynamic range video and images [2, 1].

The proposed image and video encoding methods postulate several radical departures from the commonly assumed encoding paradigms. The proposed formats encode all colors and luminance levels that are visible to the human eye, and not only the colors that can be displayed with the existing display technologies. This makes the proposed formats independent of any display or capture technologies and capable of accommodating all future technological improvements. Furthermore, the proposed formats encode visual content using scene-referred representations, which captures appearance of the actual scenes and not their rendering on a particular display, as in the case of display-referred representation. These assumptions are still not recognized by the existing image and video encoding standards, which are display-referred and limited in their precision to a particular display technology (mostly CRT displays).

Our recent project involved research and development of a backward-compatible HDR MPEG video encoding algorithm, intended for the distribution of DVD content [1]. The concept of the encoding algorithm is illustrated in Figure 15.43. The backward compatibility is achieved by encoding the HDR and low dynamic range (LDR) video frames in two video streams: an LDR stream that is compatible with MPEG decoders, and a residual stream that enables the restoration of the original HDR stream. To minimize redundancy of information, the residual and LDR streams are decorrelated. Such decorrelation requires perceptually

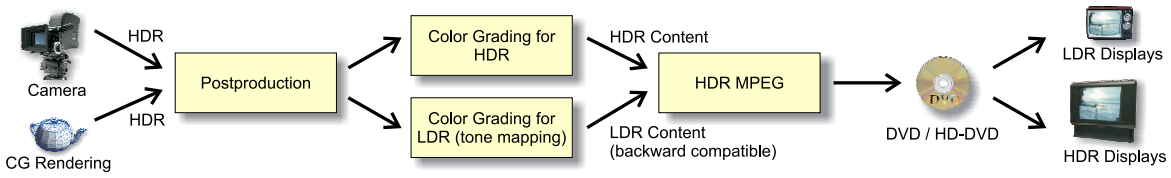


Figure 15.43: The backward-compatible HDR DVD movie processing pipeline. The high dynamic range content, provided by advanced cameras and CG rendering, is encoded in addition to the low dynamic range (LDR) content in the video stream. The files compressed with the proposed HDR MPEG method can play on existing and future HDR displays.

meaningful comparison of the LDR and HDR pixels, which we achieve by introducing a pair of corresponding color spaces that are scaled in terms of the human visual system (HVS) response to luminance and chrominance stimuli. The size of the residual stream is further reduced with a perceptually motivated filter, which removes invisible noise from video frames. To reduce the production costs of HDR DVD players, the compression algorithm is designed so that standard 8-bit MPEG decoding chip sets can be used to decode the HDR stream.

References

- [1] R. Mantiuk, A. Efremov, K. Myszkowski, and H.-P. Seidel. Backward compatible high dynamic range mpeg video compression. In J. Dorsey, ed., *Proceedings of ACM SIGGRAPH 2006*, Boston, MA, USA, July 2006, *ACM Transactions on Graphics*, vol. 25, pp. 713–723. ACM. Proc. of ACM SIGGRAPH '06.
- [2] R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Lossy compression of high dynamic range images and video. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., *Human Vision and Electronic Imaging XI*, San Jose, USA, February 2006, *SPIE*, vol. 6057, p. 60570V. SPIE.

15.12.2 Contrast-domain Image Processing

Investigators: Rafał Mantiuk and Karol Myszkowski

Since the human eye is the most sensitive for either spatial or temporal changes, we propose a framework for image operations on spatial contrast, rather than on an absolute pixel values [2, 1]. Such approach gives better control over the visibility of the changes introduced by image processing operations. The proposed framework is especially suitable for a detail-preserving compression of contrast in HDR images. Some examples of such processing are shown in Figure 15.44.

Within the proposed framework images are processed in the visual response space, in which contrast values directly correlate with their visibility in an image. Our framework involves a transformation of an image from luminance space to a pyramid of low-pass contrast images and then to the visual response space. After modifying response values, the transformation can be reversed to produce the resulting image. To predict the visibility of suprathreshold contrast, we derive a transducer function for the full range of contrast levels that can be found in high dynamic range images. We show that a complex contrast compression operation,

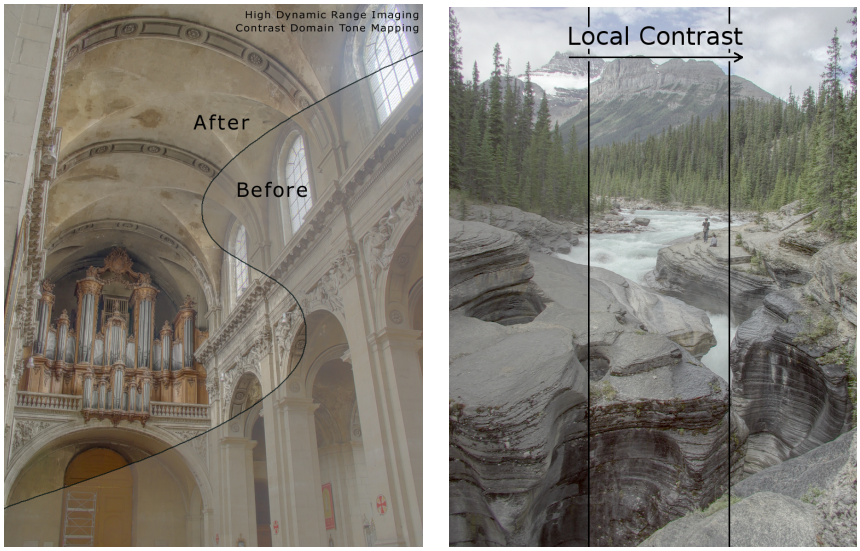


Figure 15.44: Examples of image enhancement performed in the contrast domain. Right image illustrates gradual enhancement of local contrast.

which preserves textures of small contrast, is reduced to a linear scaling in the proposed visual response space.

References

- [1] R. Mantiuk, K. Myszkowski, and H.-P. Seidel. A perceptual framework for contrast processing of high dynamic range images. In J. Malik and J. J. Koenderink, eds., *APGV '05: Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization*, Coruna, Spain, August 2005, pp. 87–94. ACM.

- [2] 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, July 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.

15.12.3 Tone Mapping Operators for HDR Images and Video

Investigators: Grzegorz Krawczyk, Kaleigh Smith, and Karol Myszkowski

The goal of this research is to address the problem of correct reproduction of HDR scene appearance during the dynamic range compression (tone mapping), limiting the distortions of the local contrasts and the change in the appearance of the original colors whenever possible.

We present a novel tone mapping operator which aims at the accurate reproduction of lightness perception of the real world scenes on low dynamic range displays [1]. The algorithm is inspired by an anchoring theory of lightness perception which comprehensively explains many characteristics of human visual system such as lightness constancy and its spectacular failures which are important in the perception of images. The principal concept of

this theory is the perception of complex scenes in terms of groups of consistent areas (frameworks). Such areas, following the gestalt theorists, are defined by the regions of common illumination. The key aspect of the image perception is the estimation of lightness within each framework through the anchoring to the luminance perceived as white, followed by the computation of the global lightness. We leverage the anchoring theory of lightness perception to handle complex images by developing an automatic method for the image decomposition into frameworks. Through the estimation of the local anchors we formalize the mapping of the luminance values to lightness. We validate the accuracy of the lightness reproduction in the presented algorithm by simulating two well known perception experiments [1, 3]. Our approach does not affect the local contrast and preserves the natural colors of an HDR image due to the linear handling of luminance. The strength of our operator is especially evident for difficult shots of real world scenes which involve distinct regions with significantly different luminance levels, Figure 15.45. Furthermore, such a decomposition into frameworks opens new grounds for local image analysis in view of human perception.

Figure 15.45: The contrast of illumination in many day-to-day situations leads to over and under exposures in the conventional photography (left). The tone mapping operator [1] reduces such a contrast by decomposing an image into the areas of consistent illumination (middle) and optimizes the contrast ratio between these areas (right).

The increasing application of HDR video demands real-time performance from tone mapping algorithms. Additionally, depiction of HDR contents with a substantial dose of realism on LDR displays requires to account for the perceptual effects which would normally appear in the real-world observation conditions but are not captured by a camera. These effects include day and night vision, temporal adaptation, change in visual acuity and glare effects. In [2] we present an efficient way to combine the most significant perceptual effects in the context of perception of HDR images including the local tone mapping into a common computational framework, which enables an efficient implementation on currently available graphics hardware. We develop a post processing module which can be added as the final stage of any real-time rendering system, game engine, or digital video player, which enhances the realism and believability of displayed HDR image streams.

For a majority of existing operators, tone mapping is achieved through the reduction of physical luminance contrast in LDR images. However, perceived image contrast is not only a function of the dynamic range of the tone mapped image, but also depends significantly on other image attributes such as lightness, hue, chroma, and sharpness. This means that by skillfully tuning these attributes, the losses in physical contrast due to tone mapping can

be restored as perceived contrast. In [4] we take a non-standard approach to the problem of depicting HDR images for LDR display. Instead of developing yet another algorithm, we provide the means to enhance the depiction of an HDR image produced by an arbitrary tone mapping algorithm, thus restoring original contrast information. The purpose of this work is two-fold: to analyze a displayed LDR image against its original HDR counterpart in terms of perceived contrast distortion, and to enhance the LDR depiction with perceptually driven color adjustments to restore the original HDR contrast information. For analysis, we present a novel algorithm for the characterization of tone mapping distortion in terms of observed loss of global contrast, and loss of contour and texture details. We classify existing tone mapping operators accordingly. We measure both distortions with perceptual metrics that enable the automatic and meaningful enhancement of LDR depictions. For image enhancement, we identify artistic and photographic color techniques from which we derive adjustments that create contrast with color. The enhanced LDR image is an improved depiction of the original HDR image with restored contrast information.

References

- [1] G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Lightness perception in tone reproduction for high dynamic range images. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, 2005, *Computer Graphics Forum*, vol. 24, pp. 635–645. Blackwell.
- [2] G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Perceptual effects in real-time tone mapping. In *SCCG '05: Proceedings of the 21st Spring Conference on Computer Graphics*, Budmerice, Slovakia, 2005, pp. 195–202. ACM.
- [3] G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Computational model of lightness perception in high dynamic range imaging. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., *Human Vision and Electronic Imaging X, IS&T/SPIE's 18th Annual Symposium on Electronic Imaging (2006)*, San Jose, CA, USA, January 2006, *SPIE*, vol. 6057, pp. 1–12. SPIE.
- [4] K. Smith, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Beyond tone mapping: Enhanced depiction of tone mapped HDR images. In L. Szirmay-Kalos and E. Gröller, eds., *The European Association for Computer Graphics 27th Annual Conference, EUROGRAPHICS 2006*, Vienna, Austria, September 2006, *Computer Graphics Forum*, vol. 25, pp. 427–438. Blackwell.

15.12.4 Psychophysical Studies of Tone Mapping Operators for Preference and Fidelity

Investigators: Akiko Yoshida, Rafał Mantiuk, Volker Blanz, and Karol Myszkowski

To visualize high dynamic range (HDR) images on low dynamic range (LDR) displays, a number of successful tone mapping operators (TMOs) inspired by image processing, photographic practice, and the characteristic of the human visual system (HVS) have been proposed. The variety of approaches calls for a systematic perceptual evaluation of their performance. We have conducted two series of psychophysical experiments to asses how human observers perceive different tone mapped images and to investigate desired properties of TMOs. Low dynamic range (LDR) and high dynamic range (HDR) displays, as well as actual scene setups, are used for both experiments.

The primary interest of the first experiment is to assess the differences in how tone mapped images which are displayed on an LDR monitor are perceived by human observers and to find out which attributes of image appearance account for these differences when tone mapped images are compared directly with their corresponding real-world scenes rather than with each other [1]. Two scenes and seven TMOs are employed, and 14 people participated. The results are analyzed by analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA). Our results demonstrate that qualitative differences in TMOs have a systematic effect on the perception of scenes. Additionally, we calculate similarities between TMOs in Mahalanobis distances based on MANOVA results. It addresses that all studied TMOs are divided into global and local categories in terms of similarity by human perception.

The goal of the second project on the perception of HDR images is to investigate the desired properties of a TMO and to design such an operator based on the data collected in a psychophysical experiment [2]. To collect this data, a series of experiments are conducted on the HDR display, in which the subjects adjust three generic TMO parameters: brightness, contrast and color saturation. We use the collected data to find a new pair of TMO parameters, which would be a more intuitive alternative to brightness and contrast adjustment. Such parameters are *reference white*, which is the luminance level in the image perceived as white, and *contrast*, which is the ratio of the lowest and the highest displayed luminance.

Another goal of this project is to find what is the subjective preference for the display brightness and dynamic range. We simulate several potential displays of different contrast and maximum brightness on the HDR display. The subjects rate the quality of reproduction of displayed images for each simulated display. Our analysis shows that people prefer brighter display of the maximum possible contrast.

References

- [1] A. Yoshida, V. Blanz, K. Myszkowski, and H.-P. Seidel. Testing tone mapping operators with human-perceived reality. *Journal of Electronic Imaging*, 16, 2007. To appear.

- [2] A. Yoshida, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Analysis of reproducing real-world appearance on displays of varying dynamic range. In *EUROGRAPHICS 2006 (EG'06)*, Vienna, Austria, September 2006, *Computer Graphics Forum*, vol. 25, pp. 415–426. Blackwell.

15.13 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 institutes benefit from software that had been developed in our group.

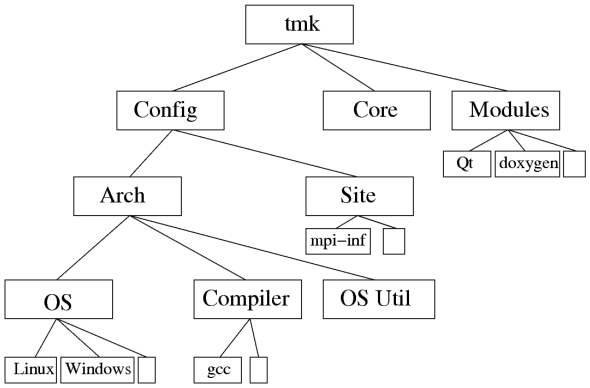


Figure 15.46: The components that build up `tmk`. The configuration is split into two major branches to separate architecture-specific configuration from site-related issues like package existence and installation paths. The CONFIG/ARCH branch is split further to distinguish between operating-system support and things like compiler environments and OS-specific helper tools.

15.13.1 TMK

Investigator: Hitoshi Yamauchi et al.

Overview

`tmk` [4] was originally developed by Hartmut Schirmacher and Stefan Brabec in 2000. `tmk` is a tool that embeds the functionality of `make` in the scripting language `tcl` [1] in a very simple and convenient way. Furthermore, `tmk` allows higher levels of abstraction via *modules* and a flexible *configuration* framework. In addition to using `tmk` simply as a replacement for `make` [5], the users can create *projects* with global methods, objects, and options, and extend or modify the globally defined tasks using per-directory control files similar to the traditional `Makefile` concept.

The design of `tmk` has been driven by the demand for two things: a simple system for managing larger software projects without having platform- or site-specific code in each `Makefile`, and a scripting environment that is combined with the core functionality of `make`. As a common basis for achieving both goals, we have chosen to embed `make`-like functions into `tcl`. Additionally, the `tmk` core natively supports architecture-dependent output, multi-directory processing, and things such as exception/exclusion handling.

On top of the `tmk` core, we have added additional abstraction layers by a module mechanism and a centralized configuration system. Through this, it is possible to remove any platform- or configuration-specific code from the control files. Figure 15.46 illustrates the hierarchy of components that build up `tmk`.

The `tmk` core

The core functionality of `tmk` was designed to embed the functionality of `make` into `tcl`. The control files for `tmk` are simply `tcl` scripts, called `TMakefile`, and define, either implicitly or explicitly, how to create a target from a set of source files (or primary dependencies).

Primary dependencies alone are used to select the appropriate rule (if there are multiple candidates) for each target, whereas secondary dependencies simply define additional preconditions before the dependent target can be built.

In addition to these most basic features, `tmk` also has a way of handling exceptions and exclusions. An *exclusion* means that some target will not be built and will not appear as dependency in any rule. An *exception* temporarily overrides the values of some variables for just some targets. Exceptions also allow to replace the rule by a completely different one.

Modules and Configuration

On top of the core, `tmk` has a *module* mechanism that allows to globally store rules, options, and procedures for certain classes of tasks. Modules are explicitly requested in the control files in order to allow the user to choose the right set of methods for the specific task, and they are parameterized through global or namespace-relative variables.

Site-dependent variables (e.g. installation paths) are not defined inside the module, but rather in the appropriate *site-config* files that are processed by `tmk`'s central *configuration system*. Similar to this, `tmk` reads *arch-config* files that define architecture-dependent options, like for example a procedure for how to call the compiler and linker for a certain task.

TMK Versions and Updates

In August 2000, the `tmk` web site [2] became online and the first beta release V0.9 was available for the public. At its initial state, `tmk` ran under IRIX, Solaris, Linux, FreeBSD as well as under Windows 98, 95, NT, and 2000, including support for many of the commonly used compilers under each operating system.

TMK Status Update – 2003

In the period 2001–2003 the basic functionality and syntax of `tmk` did not change. However, several bug fixes and improvements were done. One of the main improvements is that `tmk` now supports parallel compilation using a simple thread mechanism. For Dual-CPU machines, but also on a single processor, compilation is extremely fast. `tmk` now also runs stable on Windows 2000 supporting several compilers (Visual Studio 6/7, cygwin gcc). In 2003, `tmk` is accepted by `sourceforge.net` as an *open source* project [3].

TMK Status Update – 2005

The basis of external software libraries supported as `tmk` modules was extended. Support for 64-bit software development under Linux was added. Several modules are updated for supporting MacOS X (Darwin) environment.

TMK Status Update – 2007

The basis of `tmk` modules was further extended, e.g., to feature compilation of MATLAB programs. Other updates refer to mature 64-bit support, new compiler versions (gcc4) and new library versions (Qt4).

Although the latest update of the public version took place in September 2004, the MPII-internal version of `tmk` is being maintained and extended continuously, since it is used excessively for nearly all projects within our working group (D4). Also, several other institutes/persons choose to use `tmk` for their projects, e.g. at the graphics lab of the University of British Columbia (UBC, Vancouver, Prof. Dr. Wolfgang Heidrich).

References

[1] J. K. Ousterhout. *Tcl and the Tk Toolkit*. Addison-Wesley, 1994.

- [2] H. Schirmacher and S. Brabec. Tmk home page (moved to tmk sourceforge). `http://www.tmk-site.org`.
- [3] H. Schirmacher and S. Brabec. Tmk sourceforge. `http://sourceforge.net/projects/tmk/`.

[4] H. Schirmacher and S. Brabec. tmk - a multi-site, multi-platform system for software development. In C. Zerbst, ed., *Proc. First European Tcl/Tk User Meeting*, Technische Universität Hamburg-Harburg, 2000. Technische Universität Hamburg-Harburg.

[5] R. Stallman and R. McGrath. GNU Make. `http://www.gnu.org/software/make`.

15.13.2 D4 Shared Projects

Since the computer graphics group (D4) was founded in 1999, great efforts have been undertaken to develop well-designed libraries that can easily be combined and extended. Some of these libraries and projects are used and developed as shared projects with other sites, as for example the University of Erlangen, the University of Aachen, and the University of British Columbia (Canada).

Shared Projects System

Investigator: Hitoshi Yamauchi et al.

From a system point of view, our shared projects focus on the following points:

- hierarchical and transparent project structure
- versioning and central repository
- shared source code basis
- shared object code and executable code
- platform-independence and support for multiple platforms in parallel
- semi-automatic documentation
- easy start and minimal resource requirements for students.

The source code is maintained in a central *CVS*[1] repository that provides a number of features for tracking changes, defining versions, and merging code fragments that have been changed concurrently by multiple users. The essential parts of this code tree are checked out from the repository and compiled on the most important platforms (currently IRIX and Linux) every night by an automatic system based on `tmk` (see 15.13.1). This results in a complete and always-up-to-date *shared version* of the complete source code, libraries,

and executables. Should the compilation fail for a certain project, the system restores the previous day's state and sends error report mails to the project administrators as well as to everybody who committed changes to these projects during the last days.

If someone wishes to use a centrally maintained library, she or he simply specifies the project-relative library directory, e.g. `IBR/image/img_core`, in the local *TMakefile*. `tmk` will then search for this library in the person's project directory, and take the shared instance of the library if it is not provided by the user. This way, library locations are completely transparent to the user and need not to be specified in the project *Makefiles*. Specifically, a new user does not need to maintain any source or object code of the shared projects in her/his directory, except if she/he wants to change that code.

If a user compiles a project or application, the generated object code will always be stored in a platform- and codelevel-dependent subdirectory, such that compilation on multiple platforms will simply generate multiple subdirectories below the source code directory. `tmk` will automatically separate all platform-specific things and also use the right version of the shared code basis.

The web-based documentation of the code is generated by using *doxygen*[2], a freely available system for semi-automatic extraction of documentation information from C/C++ source code. The documentation is updated every night in the same way (and with the same methods) as it is done for the code.

References

[1] Concurrent versioning system (cvs) home page. `http:://www.cvshome.org`

[2] Doxygen home page. `http://www.doxygen.org`

BASE Project

Investigators: Christian Rössl, Hitoshi Yamauchi et al.

The D4 base library is a collection of useful general-purpose classes, functions, macros, and so on:

- vectors, matrices, and helpers
- reference counting and smart pointers
- files and filenames, path lists
- value representation as strings
- key/value parameters and parameter lists
- applications with decoupled OpenGL output
- graphical user interfaces (*Qt* widgets with extended functionality).

Single or multiple features of the base project are widely used within the group and by the projects described below. Like all our projects, *base* is organized as a number of mostly independent directories/libraries.

One of the biggest recent efforts is the porting of the source code to the most recent Qt4 libraries (by Hitoshi Yamauchi).

Image Based Rendering Code Base

Investigators: Hendrik Lensch, Michael Goesele, Christian Fuchs et al.

This IBR code base contains some basic types and functions for tasks in the context of image-based modeling and rendering. Of course the most central concept in this context is that of an *image*. An image is a collection of multiple *buffers*, each of which is a regular array of scalars or vectors of a certain type, one vector or scalar representing one layer of a *pixel* in the image. For example, an image can consist of a RGB-byte-valued *color* buffer plus a floating point-valued *depth* buffer. Additionally, every buffer may have a list of parameters (key/value pairs) that are represented and stored (I/O) transparently. Images and buffers can be stored in a proprietary and portable file format (called *IBRraw*), or they can (with some limitations) be mapped to well-known image formats such as *PPM*, *PNG*, *TIFF*, but also to specialized image formats for high dynamic range (HDR) images such as *OpenEXR* or *Radiance Picture File Format*. For further extensibility, the IBR project provides a general plug-in mechanism for I/O subroutines, the so-called *loaders* and *writers*.

Operating on images, the IBR project defines generic *filter* classes. A filter takes one or multiple images as input, and produces one or many images as output. It is parameterized via a generic string-based parameter interface (using the *base* project's key/value and parameter interface), and a graphical user interface can be generated automatically for every filter. Meanwhile a lot of different filters have been implemented for different kinds of tasks, such as standard image processing (color manipulation, arithmetic operations on pixels, cropping, resizing, blurring, etc.), reconstruction of colors from sensors (e.g. we have implemented an alternative color reconstruction filter for our Kodak DCS professional digital camera), color space conversion, type conversion (short/float/byte values, packing, coding, compression), buffer-specific routines such as conversion between different types of depth value representations, and many more.

Furthermore, the IBR project also contains routines for handling large collections of images in memory that is shared by multiple processes, and for caching images in shared memory. There is also a number of routines for mapping images to textures, generating image hierarchies, and similar things. Last, but not least, there are some command line tools as well as a graphical framework for using images and filters via their generic interfaces.

On top of these data structures and algorithmic framework, many different projects have been carried out.

Geometric Modeling Utilities

Investigators: Christian Rössl, Hitoshi Yamauchi et al.

The *Geometric Modeling Utilities (GMU)* library provides support for geometry processing with focus on handling polygonal meshes. The core of the library is a set of generic data structures for the representation and manipulation of triangle meshes. The library is written in `C++`. Current versions support the IRIX MIPSPro and the GNU compilers. *GMU* is organized in several modules:

- low-level math code (e.g. matrix computations, optimization)
- low-level system dependent code (e.g. general purpose IO, parallel processing)

- input/output of triangle meshes (multiple file formats, transparent to the user, easily expandable)
- core triangle mesh data structures
- several supporting modules for high level operations such as mesh decimation, subdivision, parameterization, filtering, discrete differential geometry
- scene graph for visualization of arbitrary data (many existing nodes e.g. for visualizing triangle meshes, easily expandable)
- user interface/widgets for rendering and exploring a scene graph (similar to *Open Inventor* widgets, based on the *Qt* library)

GMU uses the *base* library, *OpenGL* for rendering and *Qt* for user interfaces. It is entirely built by *tmk* and provides documentation and examples.

The fact that many of the new D4 members quickly adopted *GMU* for their research demonstrates its intuitiveness and efficiency best.

Starting in 2004, large parts of *GMU* including the core were completely redesigned and re-implemented. The programming interface is still very similar, however, the library is much more dynamic and less dependent on static `template` constructs.

Until 2007, several high-level features such as frameworks and geometric algorithms were added, e.g., new and configurable mesh decimation and segmentation, extensible application framework, full featured scripting (currently with bindings for `tcl` and `ruby`), a lightweight plug-in interface, and support of interaction with MATLAB (e.g., to directly access new mesh parameterization algorithms). One of the biggest recent efforts is the porting of the source code to the most recent Qt4 libraries (by Hitoshi Yamauchi).

15.13.3 PFSTOOLS for Processing High Dynamic Range Images and Video

Investigators: Rafał Mantiuk and Grzegorz Krawczyk

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 15.12. 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 XI, IS&T/SPIE's 18th Annual Symposium on Electronic Imaging (2007)*, San Jose, California, January 2007, SPIE Proceedings Series, pp. 1–12. SPIE.

15.14 Academic Activities

15.14.1 Journal Positions

Volker Blanz is on the editorial board of

- *Computer Animation and Virtual Worlds* (since 2004).

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).

15.14.2 Conference and Workshop Positions

Membership in Program Committees

Alexander Belyaev:

- *Eurographics/ACM SIGGRAPH Symposium on Geometry Processing (SGP 2007)*, Barcelona, July 2007 (Program Co-Chair),
- *IEEE International Conference on Shape Modeling and Applications (SMI 2006)*, Matsushima, June 2006 (Program Co-Chair),
- *International Conference on Shape Modeling and Applications (SMI 2005)*, MIT, Cambridge, MA, June 2005 (Program Co-Chair),
- *3rd International Symposium on Visual Computing (ISVC 2007)*, Lake Tahoe, Nevada/ California, November 2007,
- *15th Pacific Conference on Computer Graphics and Applications (PG 2007)*, Maui, Hawaii, October-November 2007,
- *10th International Conference on Computer-Aided Design and Computer Graphics (CAD/Graphics 2007)*, Beijing, October 2007,
- *Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, Belo Horizonte, Minas Gerais, October 2007,
- *New Advances in Shape Analysis and Geometric Modeling (NASAGEM 2007)*, Hannover, October 2007,
- *12th IMA conference on Mathematics of Surfaces*, Sheffield, September 2007,
- *SIGGRAPH Sketches and Posters 2007*, San Diego, California, August 2007,
- *12th ACM Symposium on Solid and Physical Modeling (SPM 2007)*, Beijing, June 2007,
- *Spring Conference on Computer Graphics (SCCG 2007)*, Budmerice Castle, April 2007,
- *Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2006)*, Manaus, Amazonas, October 2006,
- *2nd International Symposium on Visual Computing (ISVC 2006)*, Lake Tahoe, Nevada/ California, September 2006,
- *9th International Conference on Humans and Computers (HC 2006)*, Aizu-Wakamatsu, September 2006.
- *16th International Conference on Computer Graphics and Applications (GraphiCon 2006)*, Novosibirsk, July 2006,
- *Geometric Modeling and Processing (GMP 2006)*, Pittsburgh, July 2006,
- *Eurographics/ACM SIGGRAPH Symposium on Geometry Processing (SGP 2006)*, Sardinia, June 2006,
- *Spring Conference on Computer Graphics (SCCG 2006)*, Budmerice Castle, April 2006,

Volker Blanz:

- *Eurographics 2007*, Prague, September 2007,

- *ACM Symposium on Computer Animation SCA*, San Diego, August 2007,
- *IEEE Computer Vision and Pattern Recognition CVPR*, Minneapolis, June 2006,
- *MIRAGE 2007 – Computer Vision/Computer Graphics Collaboration Techniques and Applications*, INRIA Rocquencourt, March 2007,
- *ACM Symposium on Computer Animation SCA*, Vienna, September 2006,
- *Int. Conf on Pattern Recognition ICPR*, Honk Kong, August 2006,
- *IEEE Computer Vision and Pattern Recognition CVPR*, New York, June 2006,
- *European Conference on Computer Vision ECCV*, Graz, May 2006,
- *International Conference on Face and Gesture Recognition*, Southampton, April 2006,
- *IEEE International Conference on Computer Vision ICCV*, Beijing, October 2005,
- *ACM Symposium on Computer Animation SCA*, Los Angeles, July 2005,
- *IEEE Computer Vision and Pattern Recognition CVPR*, San Diego, June 2005.

Hendrik P. A. Lensch:

- *Eurographics*, Crete, April 2008,
- *Eurographics Symposium on Rendering*, Grenoble, June 2007,
- *Workshop on Photometric Analysis For Computer Vision (PACV) in conjunction with ICCV*, Rio de Janeiro, October 2007,
- *ACM SIGGRAPH*, San Diego, August 2007,
- *International Conference on Computer Graphics Theory and Application*, Barcelona, March 2007,
- *MIRAGE 2007 – Computer Vision/Computer Graphics Collaboration Techniques and Applications*, INRIA Rocquencourt, March 2007,
- *Eurographics Symposium on Rendering 2006*, Nicosia, June 2006,
- *Computer Graphics International*, Hangzhou, June 2006,
- *Graphics Interface*, Victoria, May 2005.

Karol Myszkowski:

- *Fourth 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,
- *Winter School on Computer Graphics (WSCG'07)*, Plzen, February 2007,
- *Human Vision and Electronic Imaging XII (HVEI'07)*, San Jose, January 2007,
- *International Conference on Cyberworlds 2006* Lausanne, November 2006,
- *Eurographics 2006*, Vienna, September 2006,
- *ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization 2006*, Boston, July 2006,
- *Eurographics Symposium on Rendering 2006*, Nicosia, June 2006,
- *Spring Conference on Computer Graphics (SCCG'06)*, Budmerice Castle, May 2006,
- *Winter School on Computer Graphics (WSCG'06)*, Plzen, February 2006,
- *Third International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa (Afrigraph'06)*, Cape Town, January 2006,
- *Human Vision and Electronic Imaging XI (HVEI'06)*, San Jose, January 2006,
- *International Conference on Cyberworlds 2005*, Singapore, November 2005,
- *International Workshop on Databases in Networked Information Systems (DNIS'05)*, Aizu Wakamatsu, September 2005,
- *ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization 2005*, A Coruña, August 2005,
- *Eurographics Symposium on Rendering 2005*, Konstanz, June 2005,
- *Graphicon 2005*, Novosibirsk Akademgorodok, June 2005,
- *Spring Conference on Computer Graphics (SCCG'05)*, Budmerice Castle, May 2005.

Bodo Rosenhahn

- *International Conference on Computer Vision (ICCV)*, Rio de Janeiro, October 2007, (Workshop Chair),
- *International Conference on Cyberworlds*, Hannover, October 2007,
- *MIRAGE 2007 – Computer Vision/Computer Graphics Collaboration Techniques and Applications*, INRIA Rocquencourt, March 2007,
- *Pacific Symposium on Image Vision Technology (PSIVT)*, Hsinchu, December 2006,
- *Image and Vision Computing New Zealand (IVCNZ)*, Great Barrier Island, November 2006,
- *International Workshop on Combinatorial Image Analysis (IWCIA)*, Berlin, June 2006,
- *Computer Graphics International (CGI)*, Hangzhou, June 2006,
- *European Conference on Computer Vision (ECCV)*, Graz, May 2006.

Hans-Peter Seidel:

- *Pacific Graphics 2007 (PG'07)*, Maui, Hawaii, October 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, June 2007,
- *ACM Symposium on Solid and Physical Modeling 2007 (SPM'07)*, Beijing, June 2007,
- *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, Seattle, April 2007,
- *Vision, Modeling and Visualization 2006 (VMV'06)*, Aachen, November 2006,
- *Eurographics 2006*, Vienna, September 2006,
- *ACM SIGGRAPH 2006*, Boston, August 2006,
- *ACM SIGGRAPH/Eurographics Symposium on Computer Animation 2006 (SCA'06)*, Vienna, September 2006,
- *DAGM Symposium 2006 (DAGM'06)*, Berlin, August 2006,
- *Computer Graphics International 2006 (CGI'06)*, Hangzhou, June 2006,
- *Shape Modeling International (SMI'06)*, Matsushima, June 2006,
- *Eurographics/ACM SIGGRAPH Symposium on Geometry Processing 2006 (SGP'06)*, Sardinia, June 2006,
- *International Symposium on 3D Data Processing, Visualization, and Transmission 2006 (3DPVT'06)*, Chapel Hill, June 2006,
- *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games 2006*, Redwood City, March 2006,
- *Vision, Modeling and Visualization 2005 (VMV'05)*, Erlangen, November 2005,
- *Pacific Graphics 2005 (PG'05)*, Macao, October 2005,
- *Eurographics 2005*, Dublin, September 2005,
- *DAGM Symposium 2005 (DAGM'05)*, Vienna, August 2005,
- *ACM SIGGRAPH/Eurographics Symposium on Computer Animation 2005 (SCA'05)*, Los Angeles, July 2005,
- *Eurographics/ACM SIGGRAPH Symposium on Geometry Processing 2005 (SGP'05)*, Vienna, July 2005,
- *Volume Graphics 2005 (VG'05)*, Stony Brooke, June 2005.

Christian Theobalt:

- *IEEE International Symposium on Multimedia 2007 (ISM2007)*, Taichung, December 2007,
- *MIRAGE 2007 – Computer Vision/Computer Graphics Collaboration Techniques and Applications*, INRIA Rocquencourt, March 2007,
- *2nd International Conference on Computer Graphics Theory and Applications (GRAPP'07)*, Barcelona, March 2007.

Membership in Organizing Committees

Bodo Rosenhahn

- *Dagstuhl Workshop Human Motion – Understanding, Modeling, Capture and Animation*, June 2006,
- *ICCV Workshop Human Motion – Understanding, Modeling, Capture and Animation II*, October 2007.

Hans-Peter Seidel:

- *Dagstuhl Seminar on Visual Computing - Convergence of Computer Graphics and Computer Vision*, April 2007.

15.14.3 Invited Talks and Tutorials

Alexander Belyaev:

- *Continuous Barycentric Coordinates*, Invited talk, Discrete Differential Geometry seminar, Institute of Mathematics, TU Berlin, Berlin, October 2006.
- *Non-conformal sparse low-degree implicit surfaces*, Invited talk, Lorraine Laboratory of IT Research and its Applications (LORIA), Nancy, November 2005.
- *Feature-sensitive clustering for geometric data processing*, Invited talk, Minisymposium on Feature Sensitive Mesh Processing, SIAM Conference on Geometric Design and Computing, Phoenix, AZ, November 2005.
- *Accurate detection of ridges in range data*, Invited talk, Dagstuhl Seminar on Geometric Modeling. Schloss Dagstuhl, May-June 2005.

Volker Blanz:

- *Computing Human Faces for Human Viewers: Automated Animation in Photographs and Paintings*, Invited talk, ICMI, Banff, November 2006.
- *A Learning-Based Algorithm for Dental Inlay Design*, Invited talk, European Conference on Mathematics for Industry, Madrid, July 2006.
- *Face Recognition based on a 3D Morphable Model*, Invited talk, International Conference on Face and Gesture Recognition, Southampton, April 2006.

Grzegorz Krawczyk:

- *Die nächste Generation der Digitalen Fotografie und Videotechnik*, Invited public lecture, Sonntags im Gespräch – Die Universität am Schlossplatz, Stadtverband Saarbrücken, Saarbrücken, March 2007.
- *HDR Capture: Methods and Calibration*, Invited talk, Deutsche Thomson-Brandt GmbH, Villingen, June 2006.
- *HDR Tonemapping for Images and Video*, Invited talk, HDR Photographie und 3D Visualisierung, Briese Studios, Hamburg, November 2005.

- *High Dynamic Range Techniques in Graphics: From Acquisition to Display*, Tutorial presenter, Eurographics, Dublin, August 2005.
- *Tonemapping for High Dynamic Range Video*, Invited talk, IMS-CHIPS GmbH, Stuttgart, June 2005.

Hendrik P. A. Lensch:

- *Reflektanzfelder*, Invited talk, Universität Erlangen-Nürnberg, Februar 2007.
- *Akquisition von Reflektanzfeldern*, Invited talk, TU Dresden, November 2006.
- *Dual Photography for Scene Relighting*, Invited talk, Pixar Animation Studios, Oktober 2005.
- *Dual Photography*, Invited talk, Hewlett Packard Research Labs, September 2005.
- *Properties of the Transport Matrix*, Invited talk, University of Southern California, September 2005.
- *Realistic Materials in Computer Graphics*, Tutorial presenter, ACM Siggraph, Boston, August 2006.

Rafał Mantiuk:

- *High Dynamic Range Imaging and Video Compression*, Invited talk, Mitsubishi Electric Research Laboratories, Cambridge, MA, August 2006.
- *Lossy Compression of High Dynamic Range Images and Video*, Invited talk, Deutsche Thomson-Brandt GmbH, Villingen, June 2006.
- *High Dynamic Range Imaging in Video Compression, Quality Estimation and Tone Reproduction*, Invited talk, Warsaw University of Technology, Warsaw, April 2006.
- *Research on High Dynamic Range Imaging at MPI Informatik*, Invited talk, Philips High Tech Campus, Eindhoven, March 2006.
- *High Dynamic Range Imaging: Lossy Compression, Contrast Domain Image Processing and Subjective Tone Mapping*, Invited talk, Sharp Laboratories of America, Camas, WA, January 2006.
- *HDR Imaging: Software, Video Coding and Contrast Domain Image Processing*, Invited talk, HDR Photographie und 3D Visualisierung, Briese Studios, Hamburg, November 2005.

Karol Myszkowski:

- *High Dynamic Range Video Compression*, Tutorial presenter, ACM Siggraph, Boston, August 2006.
- *Beyond Tone Mapping: Enhanced Depiction of Tone Mapped HDR Images*, Invited talk, Ayia Napa Seminar on Computer Graphics, Cyprus, June 2006.
- *High Dynamic Range Research at MPI*, Invited talk, Deutsche Thomson-Brandt GmbH, Villingen, June 2006.
- *High Dynamic Range Media*, Invited talk, 3rd Saxonian Day of Middle and East-Europe, Dresden, June 2006.

- *Beyond Tone Mapping: Enhanced Depiction of Tone Mapped HDR Images*, Invited talk, Dagstuhl Seminar on Computational Aesthetics in Graphics, Visualization and Imaging, Dagstuhl, June 2006.
- *High Dynamic Range Visible Differences Predictor*, Invited talk, HDR Photographie und 3D Visualisierung, Briese Studios, Hamburg, November 2005.
- *High Dynamic Range Techniques in Graphics: From Acquisition to Display*, Tutorial co-organizer, Eurographics, Dublin, August 2005.
- *State-of-the-art in Tone Mapping*, Invited talk, IMS-CHIPS GmbH, Stuttgart, June 2005.

Hans-Peter Seidel:

- Theory and Practice of Computer Graphics, 25th Eurographics UK Conference, Bangor, June 2007.
- 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.
- Dagstuhl Seminar on Theoretical Foundations of Computer Vision: Human Motion - Understanding, Modeling, Capture and Animation, Dagstuhl, June 2006.
- Dagstuhl Seminar on Computational Aesthetics in Graphics, Visualization and Imaging, Dagstuhl, June 2006.

Christian Theobalt:

- *Video-based Capturing and Rendering of People*, Invited talk, University of Washington, Seattle, August 2006.
- *Video-based Capturing and Rendering of People*, Invited talk, Stanford University, Palo Alto, August 2006.
- *Video-based Capturing and Rendering of People*, Invited talk, ETH Zürich, July 2006.
- *3D Fernsehen – die Computergrafik erforscht die Zukunft visueller Medien*, Invited talk, Wissenschaftssommer 2006, Munich, July 2006.
- *Video-based Capturing and Rendering of People*, Invited talk, Seminar on Human Motion: Understanding, Modeling, Capture and Animation, IBFI Schloss Dagstuhl, June 2006,
- *Video-based Rendering*, Tutorial presenter, ACM Siggraph, Los Angeles, August 2006.

15.14.4 Other Academic Activities

Karol Myszkowski:

- Review Panel for the EU-sponsored CROSSMOD project IST-014891-2 (since 2006)

Hans-Peter Seidel:

- Chair, European Association of Computer Graphics (Eurographics) (2005-2006),

- 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),
- 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).

15.15 Teaching Activities

Summer Semester 2005

Courses:

Scientific Visualization (H. Theisel)

3D Image Analysis and Synthesis (V. Blanz, M. Goesele, M. Magnor)

Winter Semester 2005/2006

Seminars:

Advanced Topics in Computer Graphics (H.-P. Seidel, P. Slusallek)

Summer Semester 2006

Courses:

Geometric Modeling (A. Belyaev, H.-P. Seidel)
Realistic Image Synthesis (P. Slusallek, K. Myszkowski)

Seminars:

Computer Graphics (B. Rosenhahn, P. Slusallek)

Diploma Theses

Carsten Fuchs: Real-Time Lighting, 2005.
Wolfram von Funck: Fully Automatic Face Detection on Triangle Meshes by Symmetry Detection 2005.
Thomas Fuchs: Computergestützte Ermittlung von Weichteildicken am menschlichen Schädel aus Computertomographie-Daten zum Zwecke der Identifizierung, 2006.
Carina Schmitt: Learning-Based Movement of Human Eyes, 2006.
Martin Sunkel: Face Detection and 3D Face Reconstruction, 2006.
Sergej Fallmann: Progressive Global Illumination for Highly Complex Models, 2007.

Lukas Heidenreich: Real-time Hierarchical Stereo Matching on Graphics Hardware, 2007.
Wenxiang Ying: Edge Based HDR Compression, 2007.

Project Classes:
Media-Threshold-Bitmap Video Stabilization (Wenxiang Ying)

15.16 Dissertations, Habilitations, Offers, Awards

15.16.1 Dissertations

Completed and Defended

-  Shin Yoshizawa: Computational Differential Geometry Tools for Surface Interrogation, Fairing, and Design, 18.12.2006.
-  Rafal Mantiuk: High-Fidelity Imaging, 14.12.2006.
-  Takehiro Tawara: Efficient Global Illumination for Dynamic Scenes, 29.11.2006.
-  Jens Vorsatz: Dynamic Remeshing and Applications, 12.06.2006.
-  Christian Theobalt: From Image-based Motion Analysis to Free-Viewport Video, 27.12.2005.
-  Irene Albrecht: Faces and Hands – Modeling and Animating Anatomical and Photorealistic Models with Regard to the Communicative Competence of Virtual Humans, 22.12.2005.
-  Christian Rössl: New Techniques for the Modeling, Processing and Visualization of Surfaces and Volumes, 20.07.2005.

Completed but not yet Defended

- Rhaleb Zayer: Numerical and Variational Aspects of Mesh Parameterization and Editing.

In Preparation

- Naveed Ahmed: Simultaneous Capture of Human Motion and Reflectance Properties from Multiple Synchronized Video Recordings.
- Boris Ajdin: Multispectral HDR Imaging
- Thomas Annen: New Algorithms for Real World Relighting.
- Tunc Aydin: High Dynamic Range Image Quality Assessment.
- Robert Bargmann: Learning-Based Speech Animation and Transfer to Novel Faces.

- Tongbo Chen: Modeling and Acquisition of Surface Appearance.
- Edilson de Aguiar: Real-Time Marker-free Model and Motion Capturing.
- Zhao Dong: Video-based Editing and Rendering.
- Christian Fuchs: Light Transport in Inhomogeneous Translucent Objects.
- Martin Fuchs: Relightable Object Representations.
- Jürgen Gall: Textured Model Based Motion Capturing.
- Johannes Günther: Dynamic Lighting Simulation.
- Torsten Langer: Discrete Differential Geometry for Shape Analysis and Modeling.
- Nils Hasler: Simulation Based Vision.
- Robert Herzog: Exploiting Coherence in Global Illumination.
- Matthias Hullin: Computational Photgraphy
- Nils Hasler: Simulation Based Vision
- Grzegorz Krawczyk: Perception-Inspired Tone Mapping.
- Torsten Langer: Discrete Differential Geometry for Shape Analysis and Modeling.
- Waqar Saleem: Shape Data Analysis and Management.
- Natascha Sauber: Multifield Visualization.
- Oliver Schall: New Techniques for Scattered Data Processing in Computer Graphics.
- Kristina Scherbaum: Learning Based Modeling and Animation of Faces.
- Thomas Schultz: Visual Analysis of DT-MRI data.
- Kuangyu Shi: Pathline Oriented Topological Visualization of Time-dependent Vector Fields.
- Kaleigh Smith: Perceptual and Artistic Techniques in Computer Graphics.
- Wenhao Song: Perception-driven Shape Interrogation and Editing.
- Carsten Stoll: Template based shape processing.
- Martin Sunkel: Image Based Generic 3D Model Adaptation.
- Wolfram von Funck: Vector Field Based Shape Deformations.
- Martin Sunkel: Image Based Generic 3D Model Adaptation
- Akiko Yoshida: Tone Mapping and Perception.
- Gernot Ziegler: GPU Usage in Advanced Video and Image Processing.

15.16.2 Offers for Faculty Positions

- V. Blanz, W2-Profesorship, Univ. Siegen, 2005.
- S. Gumhold, W3-Professorship, Univ. Dresden, 2005.
- V. Havran, Lecturer, TU Prague, 2006.
- I. Ivrissimtzis, Lecturer, Durham University UK, 2006.
- J. Kautz, Lecturer, University College London, 2005.

– M. Neff, Assistant Professor (Tenure Track), UC Davis, 2006.
– H. Theisel, W2-Professorship, Univ. Bielefeld, 2006.
– H. Theisel, W3-Professorship, Univ. Magdeburg, 2007.

15.16.3 Awards

– Martin Fuchs
  VDI-Preis, 2005
– Jürgen Gall
  SMI-DAGM Graduation Award, 2006
– Michael Goesele
  Eduard Martin Award, 2005
– Grzegorz Krawczyk
  SCCG Best Paper Award, 1st Prize, 2005
– Hendrik P. A. Lensch
  EUROGRAPHICS Young Researcher Award, 2005
– Rafał Mantiuk
  Heinz Billing Award for the Advancement of Scientific Computation 2006
  High Dynamic Range Imaging: Towards the Limits of the Human Visual Perception
– Waqar Sallem
  AIM@SHAPE Best Paper Award, 2006
– Bodo Rosenhahn
  DAGM Main Price 2005
  IVCNZ Best Student Paper Award, 2005
– Christian Theobalt
  Eduard Martin Award, 2006.
– Ingo Wald
  SaarLB Wissenschaftspreis, 2005

15.17 Grants and Cooperations

There are local cooperations within the institute and with colleagues in the computer science department. In particular, there is the ongoing research project with Philipp Slusallek's group on interactive rendering and global illumination techniques and with Joachim Weickert in the field of Visual Computing and Visualization. Within the MPI we have had common projects with the former IRG3 (Dr. Marcus Magnor) in the field of "Model-based Motion Capture and Free-Viewpoint Video" and a collaboration with D1 (Prof. Kurt Mehlhorn) in the field of "Surface Reconstruction". Furthermore scientists of our group have interdisciplinary contacts within the Saarland University at the Institut für Rechtsmedizin (Dr. Dieter Buhmann), the Institut für Phonetik (Prof. Dr. William J. Barry) and DFKI (Prof. Dr. Hans Uszkoreit and Prof. Dr. Manfred Pinkal) in the field of facial modeling and speech synchronized animation.

15.17.1 Projects funded by the European Union (EU)

The following projects are funded by the European Union (EU).

Real Time Visualization of Complex Reflectance Behavior in Virtual Prototyping (RealReflect)

RealReflect aims at developing physically correct simulation of light distribution and reflection as well as an image based real-time visualization technology for synthetic objects with complex reflectance behavior. This new technology will be integrated into an existing VR-system and tested in different application scenarios in automotive industry, like the simulation of safety and design aspects in the automotive industry as well as photorealistic VR simulations in architecture. Based on the acquisition of real reflectance properties of materials, the objective of RealReflect is to develop a novel image based physically correct visualization technology for VR systems. The overall system addresses two hot issues in Virtual Reality: photo realism through visualization of real reflectance behavior of surfaces and a sophisticated light simulation that allows for a highly accurate determination of light distribution and reflection behavior. The work on RealReflect started in April 2002 and was continued until October 2005. The project has been evaluated as highly successful by the EU commission.

Partners in the project are: TU Vienna (Austria), Univ. Bonn, (Germany), Academy of Sciences (Czech Republic), INRIA (France), MPI Informatik (Germany), Daimler Chrysler AG (Germany), IC:IDO (Germany), Faurecia (France), and Virtual Reality Architecture (Austria).

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)

The mission of the Network of Excellence (NoE) AIM@SHAPE is to advance research in the direction of semantic-based shape representations and semantic-oriented tools to acquire, build, transmit, and process shapes with their associated knowledge. The overall goal is to develop a new generation of shapes in which knowledge is explicitly represented and, therefore, can be retrieved, processed, shared, and exploited to construct new knowledge.

The AIM@SHAPE consortium pursues lasting integration both at the foundational level, by initiating a new theory of digital shapes, and at the component level, by developing a digital shape workbench as a common platform for shape models and software tools. The geometric modeling team of Dept. 4 participates actively in the AIM@SHAPE project. One of the main joint project activities consists of an advanced Digital Shape Workbench (DSW). The main components of DSW are the project shape repository (http://shapes.aim-at-shape.net, maintained by our team), tools repository (http://www-sop.inria.fr/aim-at-shape, maintained by our INRIA partners), and central web portal (maintained by our team). The primary goal of the shape repository is to provide a shared collection of standard test cases and benchmarks in order to enable efficient prototyping as well as practical evaluation on real-world and large-scale shape models. Its distinctive feature is a full documentation

of the most interesting geometric properties provided by detailed metadata of the common shape ontology. High-quality models are acquired specifically for the repository and tools from the Tool Repository are integrated to automatically extract metadata for certain shape categories.

The annual reviews confirmed AIM@SHAPE as a very successful project. The work on AIM@SHAPE started in January 2004, and the project will run until December 2007.

Partners in the project are: CNR-IMATI-GE (Italy), Università di Genova (Italy), EPFL Lausanne (Switzerland), FHG/IGD (Germany), INPG Grenoble (France), INRIA (France), ITI-CERTH (Greece), MPI Informatik (Germany), UNIGE (Switzerland), SINTEF (Norway), Technion (Israel), Utrecht University (Netherlands), and Weizmann Institute of Science (Israel).

3DTV-Integrated Three-Dimensional Television – Capture, Transmission and Display (3DTV)

The Network of Excellence (NoE) 3DTV aims at creating exact 3D moving images as ghost-like replicas of 3D objects as an ultimate goal in video science. Capturing 3D scenery, processing the captured data for transmission, and displaying the result for 3D viewing are the main functional components. These components encompass a wide range of disciplines: imaging and computer graphics, signal processing, telecommunications, electronics, optics and physics are needed. The activities of the computer graphics group are focused on capturing 3D scenes for providing the "input" to the 3DTV system. Another title for the activities of this group could very well be "3DTV camera techniques" The work on 3DTV started in September 2004. The duration is 48 months and the project will run until August 2008.

Partners in the network are: Bilkent Üniversitesi (Turkey), Bremer Institut für angewandte Strahltechnik GmbH (Germany), Bulgarian Academy of Sciences (Bulgaria), De Montfort University (United Kingdom), Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V. (Germany), FogScreen Oy (Finland), Univ. Ilmenau (Germany), Tampere University of Technology (Finland), Centre For Research and Technology Hellas (Greece), Koç University (Turkey), Middle East Technical University (Turkey), Momentum Bilgisayar Yazilim Danismanlik Ticaret A.S. (Turkey), MPI Informatik (Germany), University of West Bohemia, (Czech Republic), Univ. Hannover (Germany), TU Berlin (Germany), Univ. Tübingen (Germany), Univ. of Aberdeen (United Kingdom), and Yogurt Bilgisayar Teknolojileri Tic. Ltd. Sti. (Turkey).

15.17.2 Projects funded by BMBF

The following project is funded by the German Ministry of Education, Science, and Technology (BMBF – Bundesminister für Bildung und Forschung).

Max Planck Center for Visual Computing and Communication

The Max Planck Center for Visual Computing and Communication (MPC-VCC) was established jointly by MPII and Stanford University in October 2003. The proposed collaboration has two intertwined goals:

- Establish a joint research program in the key information technology area of "Visual Computing and Communication".

- Incorporate a strong career development component to alleviate the shortage of qualified faculty and scientists in information technology in Germany.

To achieve these goals, Stanford University and the Max Planck Society collaborate to set up a Max Planck Center for Visual Computing and Communication, with corresponding research activities at the Max Planck Institute for Computer Science and at Stanford University. The center is directed jointly by Prof. H.-P. Seidel (MPII) and Prof. Dr. B. Girod (Stanford University).

The Max Planck Center for Visual Computing and Communication addresses the particular career-development needs of young German scientists in Information Technology. It 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 a senior researcher at the Max Planck Institute for Computer Science and ultimately as a professor at a German university.

15.17.3 Cooperations with Industry

The following projects have been funded by industry.

Daimler Chrysler: Simulation of Displays Appearance in the Car Cockpit

The goal of this project is to simulate the reflection of light in the surface of displays, which are installed in modern cars (navigation panels and speedometer panels) for various lighting conditions. The lighting simulation is physically-based and the computation is performed at interactive speeds. Based on the simulation results the visibility of displayed information assessed for typical and extreme lighting conditions taking into account the human perception. Since active displays are a source of lighting energy on their own, investigations is performed to estimate how the display presence affects the visual performance of the driver. In particular, reflections of the displays in the windshield is studied to assess their influence on the driving security. As a result of the project a virtual reality system running in the CAVE installed in the Virtual Reality Center has been developed. The project started in February 2003 and has been successfully completed in August 2006.

BrightSide Technologies (Dolby): Backward Compatible High Dynamic Range Video Compression

The goal of this project was research and development of a backward-compatible video compression algorithm for high dynamic range content. Such format is intended for the future generation of the DVD format, which is compatible both with standard video players

and displays as well as the next generation high dynamic range displays that offer much higher brightness, contrast and wider color gamut.

The developed video compression algorithm encodes standard dynamic range as well as high dynamic range content in a single video stream. To achieve high efficiency, both video sequences are decorrelated and filtered using a perceptual filter that removes invisible noise. The proposed compression method does not impose restrictions and does not modify the appearance of both video streams. Further details can be found in Section 15.12.1.

15.18 Publications

Journal articles and book chapters

[1] I. Albrecht, M. Schröder, J. Haber, and H.-P. Seidel. Mixed feelings: Expression of non-basic emotions in a muscle-based talking head. *Virtual Reality*, 8(4):201–212, September 2005.

[2] G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, and T. Malzbender, eds. *Advances in Visual Computing, Second International Symposium, ISVC 2006, Part I*, Berlin, Germany, 2006, *LNCS 4291*. Springer.

[3] G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, and T. Malzbender, eds. *Advances in Visual Computing, Second International Symposium, ISVC 2006, Part II*, Berlin, Germany, 2006, *LNCS 4292*. Springer.

[4] S. Brabec, T. Annen, and H.-P. Seidel. Practical shadow mapping. In R. Barzel, ed., *Graphics Tools: The JGT Editors' Choice*, pp. 217–228. A K Peters, Wellesley, MA, USA, August 2005.

[5] J. Estrada, D. Martínez, D. Leon, and H. Theisel. Solving geometric problems using subdivision methods and range analysis. In M. Daehlen, K. Morken, and L. Schumaker, eds., *Mathematical Methods for Curves and Surfaces: Tromso 2004*, pp. 101–114. Nashboro Press, Brentwood, USA, 2005.

[6] M. Fuchs, V. Blanz, H. P. A. Lensch, and H.-P. Seidel. Reflectance from images: A model-based approach for human faces. *IEEE Transactions on Visualization and Computer Graphics*, 11(3):296–305, May 2005.

[7] J. Gall, J. Potthoff, C. Schnoerr, B. Rosenhahn, and H.-P. Seidel. Interacting and annealing particle filters: Mathematics and a recipe for applications. *Journal of Mathematical Imaging and Vision*, X, 2007.

[8] 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*. Springer, Heidelberg, 2007.

[9] S. Gumhold. Truly selective polygonal mesh hierarchies with error control. *Computer Aided Geometric Design*, 22(5):424–443, 2005.

[10] J. Haber, M. Magnor, and H.-P. Seidel. Physically based simulation of twilight phenomena. *Transactions on Graphics*, 24(4):1353–1373, October 2005.

[11] I. Ivrissimtzis and D. Singerman. Regular maps and principal congruence subgroups of hecke groups. *European Journal of Combinatorics*, 26(3-4):437–456, April 2005.

[12] I. Ivrissimtzis, R. Zayer, and H.-P. Seidel. Polygonal decompositions of quadrilateral subdivision meshes. *Computer Graphics & Geometry*, 7(1):16–30, 2005.

[13] F. Jiang, V. Blanz, and A. J. O'Toole. Probing the visual representation of faces with adaption. a view from the other side of the mean. *Psychological Science*, 17(6):493–500, 2006.

[14] F. Jiang, V. Blanz, and A. J. O'Toole. The role of familiarity in three-dimensional view transferability of face identity adaptation. *Vision Research*, to appear, 2007.

[15] X. Jiang, E. Rosen, T. Zeffiro, J. VanMeter, V. Blanz, and M. Riesenhuber. Evaluation of a shape-based model of human face dicrimination using fmri and behavorial techniques. *Neuron*, 50:159–172, 2006.

[16] T. Langer, A. Belyaev, and H.-P. Seidel. Exact and interpolatory quadratures for curvature tensor estimation. *Computer Aided Geometric Design*, Accepted, 2007.

[17] Y. Lee, S. Lee, A. Shamir, D. Cohen-Or, and H.-P. Seidel. Mesh scissoring with minima rule and part salience. In L. Kobbelt, ed., *Geometry processing*, *Computer Aided Geometric Design*, vol. 22, pp. 444–465. Elsevier, Amsterdam, The Netherlands, 2005.

[18] Y. Lipman, O. Sorkine, M. Alexa, D. Cohen-Or, D. Levin, C. Rössl, and H.-P. Seidel. Laplacian framework for interactive mesh editing. *International Journal of Shape Modeling*, 11(1):43–61, 2005.

[19] 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, July 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.

[20] A. Mehl and V. Blanz. A new approach for automatic reconstruction of occlusal surfaces with the biogeneric tooth model. *Int J Comput Dent*, 8:13–25, 2005.

[21] A. Mehl, V. Blanz, and R. Hickel. Was ist der "Durchschnittszahn"? - Ein mathematisches Verfahren für die automatische Berechnung einer repräsentativen Kaufläche. *Deutsche Zahnärztliche Zeitschrift*, 60(6):335–341, 2005.

[22] M. Neef, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics*, 26, 2007.

[23] G. Nürnberger, C. Rössl, F. Zeilfelder, and H.-P. Seidel. Quasi-interpolation by quadratic piecewise polynomials in three variables. *Computer Aided Geometric Design*, 22:221–249, 2005.

[24] Y. Ohtake, A. Belyaev, and H.-P. Seidel. 3d scattered data interpolation and approximation with multilevel compactly supported rbfs. *Graphical Models*, 67(3):150–165, 2005.

[25] Y. Ohtake, A. Belyaev, and H.-P. Seidel. Multi-scale and adaptive cs-rbfs for shape reconstruction from cloud of points. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, eds., *Advances in Multiresolution for Geometric Modelling*, pp. 143–154. Springer, Berlin, Germany, 2005.

[26] Y. Ohtake, A. Belyaev, and H.-P. Seidel. A composite approach to meshing scattered data. *Graphical Models*, 68(3):255–267, 2006.

[27] Y. Ohtake, A. Belyaev, and H.-P. Seidel. Sparse surface reconstruction with adaptive partition of unity and radial basis functions. *Graphical Models*, 68(1):15–24, January 2006.

[28] S. Romdhani, V. Blanz, C. Basso, and T. Vetter. Morphable models of faces. In S. Z. Li and A. K. Jain, eds., *Handbook of Face Recognition*, ch. 10, pp. 217–245. Springer, New York, USA, 2005.

[29] B. Rosenhahn, T. Brox, U. Kersting, A. Smith, J. Gurney, and R. Klette. A system for marker-less motion capture. *Künstliche Intelligenz*, 20(1):45–51, January 2006.

[30] B. Rosenhahn, U. Kersting, K. Powell, R. Klette, G. Klette, and H.-P. Seidel. A system for articulated tracking incorporating a clothing model. *Machine Vision and Applications*, 18(1):25–40, 2007.

[31] B. Rosenhahn, C. Perwass, and G. Sommer. Pose estimation of free-form contours. *International Journal of Computer Vision*, 62(3):267–289, May 2005.

[32] B. Rosenhahn and G. Sommer. Pose estimation in conformal geometric algebra. part i: The stratification of mathematical spaces. *Journal of Mathematical Imaging and Vision*, 22(1):27–48, January 2005.

[33] B. Rosenhahn and G. Sommer. Pose estimation in conformal geometric algebra. part ii: Real-time pose estimation using extended feature concepts. *Journal of Mathematical Imaging and Vision*, 22(1):49–70, January 2005.

[34] W. Saleem, O. Schall, G. Patanè, A. Belyaev, and H.-P. Seidel. On stochastic methods for surface reconstruction. *The Visual Computer*, XX, 2007. to appear.

[35] O. Schall, A. Belyaev, and H.-P. Seidel. Error-guided adaptive fourier-based surface reconstruction. *Computer-Aided Design*, XX, 2007. to appear.

[36] P. Sen, B. Chen, G. G. Garg, S. Marschner, M. Horowitz, M. Levoy, and H. P. A. Lensch. Dual photography. *ACM Trans. on Graphics (Proc. SIGGRAPH 2005)*, 24(3):745–755, August 2005.

[37] M. Tarini, H. P. A. Lensch, M. Goesele, and H.-P. Seidel. 3d acquisition of mirroring objects. *Graphical Models*, 67(4):233–259, July 2005.

[38] H. Theisel, T. Weinkauf, H.-C. Hege, and H.-P. Seidel. Topological methods for 2d time-dependent vector fields based on stream lines and path lines. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):383–394, May 2005.

[39] C. Theobalt, N. Ahmed, H. 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(3), May 2007.

[40] I. Wald, H. Friedrich, G. Marmitt, P. Slusallek, and H.-P. Seidel. Faster isosurface ray tracing using implicit kd-trees. *IEEE Transactions on Visualization and Computer Graphics*, 11(5):562–572, September 2005.

[41] H. Yamauchi, S. Gumhold, R. Zayer, and H.-P. Seidel. Mesh segmentation driven by gaussian curvature. *The Visual Computer*, 21(8-10):649–658, September 2005.

[42] H. Yamauchi, H. P. A. Lensch, J. Haber, and H.-P. Seidel. Textures revisited. *The Visual Computer*, 21(4):217–241, May 2005.

[43] M. Yoon, Y. Lee, S. Lee, I. Ivrissimtzis, and H.-P. Seidel. Surface and normal ensembles for surface reconstruction. *Computer-Aided Design*, XX, 2007.

[44] A. Yoshida, V. Blanz, K. Myszkowski, and H.-P. Seidel. Testing tone mapping operators with human-perceived reality. *Journal of Electronic Imaging*, 16, 2007. To appear.

[45] S. Yoshizawa, A. Belyaev, and H.-P. Seidel. A moving mesh approach to stretch-minimizing mesh parameterization. *International Journal of Shape Modeling*, 11(1):25–42, June 2005.

Conference articles

[1] E. de Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel. Reconstructing human shape and motion from multi-view video. In *2nd European Conference on Visual Media Production (CVMP)*, London, UK, December 2005, pp. 42–49. The IEE.

[2] E. de Aguiar, C. Theobalt, and H.-P. Seidel. Automatic learning of articulated skeletons from 3D marker trajectories. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, and T. Malzbender, eds., *Advances in Visual Computing, Second International Symposium, ISVC 2006, Part I*, Lake Tahoe, NV, USA, November 2006, *LNCS 4291*, pp. 485–494. Springer.

[3] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less deformable mesh tracking for human shape and motion capture. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, USA, June 2007. IEEE.

[4] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Rapid animation of laser-scanned humans. In *IEEE Virtual Reality 2007*, Charlotte, USA, 2007, pp. 223–226. IEEE.

[5] E. de Aguiar, R. Zayer, C. Theobalt, M. Magnor, and H.-P. Seidel. Video-driven animation of human body scans. In *IEEE 3DTV Conference*, Kos Island, Greece, 2007. IEEE.

[6] N. Ahmed, E. de Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel. Automatic generation of personalized human avatars from multi-view video. In *VRST '05: Proceedings of the ACM symposium on Virtual reality software and technology*, Monterey, USA, December 2005, pp. 257–260. ACM.

[7] N. Ahmed, C. Theobalt, and H.-P. Seidel. Spatio-temporal reflectance sharing for relightable 3d video. In A. Gagalowicz and W. Philips, eds., *Third International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, MIRAGE 2007*, INRIA Rocquencourt, France, March 2007, *LNCS 4418*, pp. 47–58. Springer.

[8] T. Annen, W. Matusik, M. Zwicker, H. Pfister, and H.-P. Seidel. Distributed rendering for multiview parallax displays. In A. J. Woods, N. A. Dodgson, J. O. Merritt, M. T. Bolas, and I. E. McDowall, eds., *Proceedings of Stereoscopic Displays and Virtual Reality Systems XIII*, San Jose, USA, 2006, *SPIE*, vol. 6055, pp. 231–240. SPIE.

[9] R. Bargmann, V. Blanz, and H.-P. Seidel. Learning-based facial rearticulation using streams of 3D scans. In B.-Y. Chen, ed., *The 14th Pacific Conference on Computer Graphics and Applications*, Taipei, Taiwan, October 2006, pp. 232–241. National Taiwan University.

[10] A. Belyaev. On transfinite barycentric coordinates. In A. Sheffer and K. Polthier, eds., *SGP 2006, Fourth Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, Cagliari, Sardinia, Italy, 2006, pp. 89–99. Eurographics.

[11] A. Belyaev and E. Anoshkina. Detection of surface creases in range data. In R. Martin, H. Bez, and M. Sabin, eds., *Mathematics of surfaces XI, 11th IMA International Conference*, Loughborough, UK, 2005, *LNCS 3604*, pp. 50–61. Springer.

[12] V. Blanz, I. Albrecht, J. Haber, and H.-P. Seidel. Creating face models from vague mental images. In L. Szirmay-Kalos and E. Gröller, eds., *EUROGRAPHICS 2006 (EG'06)*, Vienna, Austria, September 2006, *Computer Graphics Forum*, vol. 25, pp. 645–654. Blackwell.

[13] V. Blanz, P. Grother, J. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, CA, USA, 2005, pp. 446–453. IEEE.

[14] M. Botsch, M. Pauly, C. Rössl, S. Bischoff, and L. Kobbelt. Geometric modeling based on triangle meshes. In *Eurographics Tutorial Notes*, Vienna, Austria, September 2006, p. 180. Eurographics.

[15] M. Botsch, M. Pauly, C. Rössl, S. Bischoff, and L. Kobbelt. Geometric modeling based on triangle meshes. In *SIGGRAPH Course Notes*, Boston, USA, 2006. ACM.

[16] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. High accuracy optical flow serves 3-D pose tracking: Exploiting contour and flow based constraints. In A. Leonarids, H. Bishof, and A. Prinz, eds., *Computer vision - ECCV 2006, 9th European Conference on Computer Vision; Part II*, Graz, Austria, 2006, *LNCS 3952*, pp. 98–111. Springer.

[17] T. Brox, B. Rosenhahn, U. Kersting, and D. Cremers. Nonparametric density estimation for human pose tracking. In K. Franke, K.-R. Müller, B. Nickolay, and R. Schäfer, eds., *Pattern Recognition, 28th DAGM Symposium*, Berlin, Germany, 2006, *LNCS 4174*, pp. 546–555. Springer.

[18] B. Chen and H. P. A. Lensch. Light source interpolation for sparsely sampled reflectance fields. In G. Greiner, J. Hornegger, H. Niemann, and M. Stamminger, eds., *Vision, Modeling, and Visualization 2005 (VMV'05)*, Erlangen, Germany, November 2005, pp. 461–469. Aka.

[19] T. Chen, M. Goesele, and H.-P. Seidel. Mesostructure from specularity. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, New York, NY, USA, 2006, vol. 2, pp. 1825–1832. IEEE.

[20] K. Dalal, A. Klein, Y. Liu, and K. Smith. A spectral approach to npr packing. In *Proceedings of the 4th International Symposium on Non-Photorealistic Animation and Rendering 2006 (NPAR'06)*, Annecy, France, 2006, pp. 71–78. ACM.

[21] A. Efremov, V. Havran, and H.-P. Seidel. Robust and numerically stable Bézier clipping method for ray tracing NURBS surfaces. In *SCCG '05: Proceedings of the 21st Spring Conference on Computer Graphics*, Budmerice, Slovakia, May 2005, pp. 127–135. ACM.

[22] F. Enders, N. Sauber, D. Merhof, P. Hastreiter, C. Nimsky, and M. Stamminger. Visualization of white matter tracts with wrapped streamlines. In C. T. Silva, E. Gröller, and H. Rushmeier, eds., *IEEE Visualization 2005 (VIS 2005)*, Minneapolis, USA, October 2005, pp. 51–58. IEEE.

[23] H. Friedrich, J. Günther, A. Dietrich, M. Scherbaum, H.-P. Seidel, and P. Slusallek. Exploring the use of ray tracing for future games. In *ACM SIGGRAPH Video Game Symposium sandbox '06: Proceedings of the 2006 ACM SIGGRAPH symposium on Videogames*, Boston, MA, USA, July 2006, pp. 41–50. ACM.

[24] C. Fuchs, T. Chen, M. Goesele, H. Theisel, and H.-P. Seidel. Volumetric density capture from a single image. In T. Möller, R. Machiraju, M.-S. Chen, and T. Ertl, eds., *Volume Graphics 2006, Eurographics / IEEE VGTC Workshop Proceedings*, Boston, USA, 2006, pp. 17–22. Eurographics.

[25] C. Fuchs, M. Goesele, T. Chen, and H.-P. Seidel. An empirical model for heterogeneous translucent objects. In J. Buhler, ed., *SIGGRAPH 2005 Sketches*, Los Angeles, USA, 2005, pp. 1–1. ACM SIGGRAPH.

[26] M. Fuchs, V. Blanz, and H.-P. Seidel. Bayesian relighting. In O. Deussen, A. Keller, K. Bala, P. Dutré, D. W. Fellner, and S. N. Spencer, eds., *Rendering Techniques 2005: Eurographics Symposium on Rendering*, Konstanz, Germany, July 2005, Rendering Techniques, pp. 157–164. Eurographics.

[27] W. von Funck, H. Theisel, and H.-P. Seidel. Shape matching based on fully automatic face detection on triangular meshes. In T. Nishita, Q. Peng, and H.-P. Seidel, eds., *Advances in Computer Graphics, 24th Computer Graphics International Conference, CGI 2006*, Hangzhou, China, 2006, *LNCS 4035*, pp. 242–253. Springer.

[28] W. von Funck, H. Theisel, and H.-P. Seidel. Vector field based shape deformations. In J. Dorsey, ed., *Proceedings of ACM SIGGRAPH 2006*, Boston, MA, USA, July 2006, *ACM Transactions on Graphics*, vol. 25, pp. 1118–1125. ACM. Proc. of ACM SIGGRAPH '06.

[29] I. Galic, J. Weickert, M. Welk, A. Bruhn, A. Belyaev, and H.-P. Seidel. Towards PDE-based image compression. In N. Paragios, O. Faugeras, T. Chan, and C. Schnoerr, eds., *Variational, geometric, and level set methods in computer vision, Third International Workshop, VLSM 2005*, Beijing, China, 2005, *LNCS 3752*, pp. 37–48. Springer.

[30] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Learning for multi-view 3D tracking in the context of particle filters. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, and T. Malzbender, eds., *Advances in Visual Computing, Second International Symposium, ISVC 2006, Part II*, Lake Tahoe, NV, USA, 2006, *LNCS 4292*, pp. 59–69. Springer.

[31] J. Gall, B. Rosenhahn, and H.-P. Seidel. Robust pose estimation with 3D textured models. In L.-W. Chang and W.-N. Lie, eds., *Advances in Image and Video Technology, First Pacific Rim Symposium, PSIVT 2006*, Hsinchu, Taiwan, December 2006, *LNCS 4319*, pp. 84–95. Springer.

[32] G. G. Garg, E.-V. Talvala, M. Levoy, and H. P. A. Lensch. Symmetric photography: Exploiting data-sparseness in reflectance fields. In T. Anenine-Möller and W. Heidrich, eds., *Rendering Techniques 2006: Eurographics Symposium on Rendering*, Nicosia, Cyprus, June 2006, pp. 251–262. Eurographics Association.

[33] M. Greeff, J. Haber, and H.-P. Seidel. Nailing and pinning: Adding constraints to inverse kinematics. In V. Skala, ed., *The 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2005 in co-operation with EUROGRAPHICS W S C G ' 2005*, Plzen, Czech Republic, February 2005, Short Paper Proceedings, pp. 125–128. UNION Agency.

[34] S. Gumhold. Optimizing markov models with applications to triangular connectivity coding. In *Proceedings of the sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-05)*, Vancouver, Canada, January 2005, pp. 331–338. SIAM.

[35] S. Gumhold, Z. Karni, M. Isenburg, and H.-P. Seidel. Predictive point-cloud compression. In *Proceedings of the Sixth Israel-Korea Bi-National Conference*, Haifa, Israel, 2005, pp. 125–129. Technion.

[36] S. Gumhold, Z. Karni, M. Isenburg, and H.-P. Seidel. Predictive point-cloud compression. In *SIGGRAPH 2005 Technical Sketches*, Los Angeles, USA, 2005. ACM.

[37] J. Günther, T. Chen, M. Goesele, I. Wald, and H.-P. Seidel. Efficient acquisition and realistic rendering of car paint. In G. Greiner, J. Hornegger, H. Niemann, and M. Stamminger, eds., *Vision, Modeling, and Visualization 2005 (VMV'05)*, Erlangen, Germany, November 2005, pp. 487–494. Aka.

[38] J. Günther, H. Friedrich, H.-P. Seidel, and P. Slusallek. Interactive ray tracing of skinned animations. *The Visual Computer*, 22(9-11):785–792, September 2006.

[39] J. Günther, H. Friedrich, I. Wald, H.-P. Seidel, and P. Slusallek. Ray tracing animated scenes using motion decomposition. *Computer Graphics Forum*, 25(3):517–525, September 2006.

[40] N. Hasler, M. Asbach, B. Rosenhahn, J.-R. Ohm, and H.-P. Seidel. Physically based tracking of cloth. In L. Kobbelt, T. Kuhlen, T. Aach, and R. Westermann, eds., *11th International Fall Workshop on Vision, Modeling, and Visualization 2006 (VMV 2006)*, Aachen, Germany, November 2006, pp. 49–56. IOS.

[41] N. Hasler, B. Rosenhahn, M. Asbach, J.-R. Ohm, and H.-P. Seidel. An analysis-by-synthesis approach to tracking of textiles. In *IEEE Workshop on Motion and Video Computing (WMVC 2007)*, Austin, USA, February 2007. IEEE Computer Society.

[42] N. Hasler, B. Rosenhahn, and H.-P. Seidel. Reverse engineering garments. In A. Gagalowicz and W. Philips, eds., *MIRAGE 2007*, Rocquencourt, France, March 2007, *LNCS 4418*, pp. 200–211. Springer.

[43] V. Havran, J. Bittner, R. Herzog, and H.-P. Seidel. Ray maps for global illumination. In O. Deussen, A. Keller, K. Bala, P. Dutré, D. W. Fellner, and S. N. Spencer, eds., *Rendering Techniques 2005: Eurographics Symposium on Rendering*, Konstanz, Germany, June 2005, pp. 43–54,311. Eurographics.

[44] V. Havran, R. Herzog, and H.-P. Seidel. Fast final gathering via reverse photon mapping. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, August 2005, *Computer Graphics Forum*, vol. 24, pp. 323–333. Blackwell.

[45] V. Havran, R. Herzog, and H.-P. Seidel. On the fast construction of spatial hierarchies for ray tracing. In I. Wald and S. G. Parker, eds., *Proceedings of the 2006 IEEE Symposium on Interactive Ray Tracing*, Salt Lake City, UT, USA, September 2006, pp. 71–80. IEEE.

[46] V. Havran, A. Neummann, G. Zotti, W. Purgathofer, and H.-P. Seidel. On cross-validation and resampling of brdf data measurements. In *SCCG '05: Proceedings of the 21st Spring Conference on Computer Graphics*, Budmerice, Slovakia, May 2005, pp. 161–168. ACM.

[47] V. Havran, M. Smyk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Interactive system for dynamic scene lighting using captured video environment maps. In O. Deussen, A. Keller, K. Bala, P. Dutré, D. W. Fellner, and S. N. Spencer, eds., *Rendering Techniques 2005: Eurographics Symposium on Rendering*, Konstanz, Germany, June 2005, pp. 31–42,311. Eurographics.

[48] M. Isenburg, I. Ivrissimtzis, S. Gumhold, and H.-P. Seidel. Geometry prediction for high degree polygons. In B. Jüttler, ed., *21st Spring Conference on Computer Graphics (SCCG 2005)*, Budmerice, Slovakia, 2005, pp. 147–152. Comenius University.

[49] F. Isgro, F. Odone, W. Saleem, and O. Schall. Clustering for surface reconstruction. In B. Falcidieno and N. Magnenat-Thalmann, eds., *1st International Workshop towards Semantic Virtual Environments*, Villars, Switzerland, 2005, pp. 156–162. MIRALab.

[50] I. Ivrissimtzis, W.-K. Jeong, S. Lee, Y. Lee, and H.-P. Seidel. Surface reconstruction with neural meshes. In *6th International Conference on Mathematical Methods for Curves and Surfaces*, Tromso, Norway, 2005, pp. 223–242. Nashboro Press.

[51] J.-R. Jiménez, K. Myszkowski, and X. Pueyo. Interactive global illumination in dynamic participating media using selective photon tracing. In *SCCG '05: Proceedings of the 21st Spring Conference on Computer Graphics*, Budmerice, Slovakia, 2005, pp. 211–218. ACM.

[52] G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Lightness perception in tone reproduction for high dynamic range images. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, 2005, *Computer Graphics Forum*, vol. 24, pp. 635–645. Blackwell.

[53] G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Perceptual effects in real-time tone mapping. In *SCCG '05: Proceedings of the 21st Spring Conference on Computer Graphics*, Budmerice, Slovakia, 2005, pp. 195–202. ACM.

[54] G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Computational model of lightness perception in high dynamic range imaging. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., *Human Vision and Electronic Imaging X, IS&T/SPIE's 18th Annual Symposium on Electronic Imaging (2006)*, San Jose, CA, USA, January 2006, *SPIE*, vol. 6057, pp. 1–12. SPIE.

[55] T. Langer, A. Belyaev, and H.-P. Seidel. Asymptotic analysis of discrete normals and curvatures of polylines. In *SCCG '05: Proceedings of the 21st spring conference on Computer graphics*, Budmerice, Slovakia, May 2005, pp. 229–232. ACM.

[56] T. Langer, A. Belyaev, and H.-P. Seidel. Exact and approximate quadratures for curvature tensor estimation. In M. Desbrun and H. Pottmann, eds., *Third Eurographics Symposium on Geometry Processing (Poster Proceedings)*, Aire-la-Ville, Switzerland, July 2005, pp. 14–15. Eurographics.

[57] T. Langer, A. Belyaev, and H.-P. Seidel. Exact and approximate quadratures for curvature tensor estimation. In G. Greiner, J. Hornegger, H. Niemann, and M. Stamminger, eds., *Vision, Modeling, and Visualization 2005 (VMV'05)*, Erlangen, Germany, November 2005, pp. 421–428. Aka.

[58] T. Langer, A. Belyaev, and H.-P. Seidel. Spherical barycentric coordinates. In D. W. Fellner, S. N. Spencer, A. Sheffer, and K. Polthier, eds., *SGP 2006, Fourth Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, Cagliari, Sardinia, Italy, June 2006, pp. 81–88. Eurographics.

[59] T. Langer, A. Belyaev, and H.-P. Seidel. Mean value coordinates for arbitrary spherical polygons and polyhedra in r^3. In *Proceedings of the Sixth AFA Conference on Curves and Surfaces*, Avignon, 2007. Accepted.

[60] E. Learned-Miller, Q. Lu, A. Paisley, P. Trainer, V. Blanz, K. Dedden, and R. E. Miller. Detecting acromegaly: Screening for disease with a morphable model. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Copenhagen, Denmark, 2006, *LNCS 4191*.

[61] Y. Lee, S. Lee, I. Ivrissimtzis, and H.-P. Seidel. Overfitting control for surface reconstruction. In D. W. Fellner, S. N. Spencer, A. Sheffer, and K. Polthier, eds., *SGP 2006: Fourth Eurographics Symposium on Geometry Processing*, Cagliari, Sardinia, Italy, June 2006, pp. 231–234. Eurographics.

[62] Y. Lee, M. Yoon, S. Lee, I. Ivrissimtzis, and H.-P. Seidel. Ensembles for surface reconstruction. In C. Gotsman, D. Manocha, and E. Wu, eds., *Proceedings of the 13th Pacific Conference on Computer Graphics and Applications*, Macao, October 2005, pp. 125–127. Welfare Printing Ltd.

[63] Y. Lee, M. Yoon, S. Lee, I. Ivrissimtzis, and H.-P. Seidel. Ensembles for normal and surface reconstructions. In *Geometric Modeling and Processing (GMP 2006)*, Pittsburgh, Pennsylvania, USA, 2006, *LNCS 4077*, pp. 17–33. Springer.

[64] G. Liu, R. Klette, and B. Rosenhahn. Collinearity and coplanarity constraints for structure from motion. In L.-W. Chang and W.-N. Lie, eds., *Advances in Image and Video Technology, First Pacific Rim Symposium, PSIVT 2006*, Hsinchu, Taiwan, December 2006, *LNCS 4319*, pp. 13–23. Springer.

[65] M. Magnor, M. Pollefeys, W. Matusik, G. Cheung, and C. Theobalt. Video-based rendering. In *ACM SIGGRAPH 2005 Course Notes*, Los Angeles, USA, July 2005. ACM SIGGRAPH.

[66] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel. Predicting visible differences in high dynamic range images - model and its calibration. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., *Human Vision and Electronic Imaging X, IS&T/SPIE's 17th Annual Symposium on Electronic Imaging (2005)*, San Jose, California USA, January 2005, *SPIE Proceedings Series*, vol. 5666, pp. 204–214. SPIE.

[67] R. Mantiuk, A. Efremov, K. Myszkowski, and H.-P. Seidel. Backward compatible high dynamic range mpeg video compression. In J. Dorsey, ed., *Proceedings of ACM SIGGRAPH 2006*, Boston, MA, USA, July 2006, *ACM Transactions on Graphics*, vol. 25, pp. 713–723. ACM. Proc. of ACM SIGGRAPH '06.

[68] 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 XI, IS&T/SPIE's 18th Annual Symposium on Electronic Imaging (2007)*, San Jose, California, January 2007, SPIE Proceedings Series, pp. 1–12. SPIE.

[69] R. Mantiuk, K. Myszkowski, and H.-P. Seidel. A perceptual framework for contrast processing of high dynamic range images. In J. Malik and J. J. Koenderink, eds., *APGV '05: Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization*, Coruna, Spain, August 2005, pp. 87–94. ACM.

[70] R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Lossy compression of high dynamic range images and video. In B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., *Human Vision and Electronic Imaging XI*, San Jose, USA, February 2006, *SPIE*, vol. 6057, p. 60570V. SPIE.

[71] M. Neff and H.-P. Seidel. Modeling relaxed hand shape for character animation. In F. J. Perales and R. B. Fisher, eds., *Articulated motion and deformable objects, 4th International Conference, AMDO 2006*, Port d'Andratx, Spain, 2006, *LNCS 4069*, pp. 262–270. Springer.

[72] Y. Ohtake, A. Belyaev, and M. Alexa. Sparse low-degree implicit surfaces with applications to high quality rendering, feature extraction, and smoothing. In M. Desbrun and H. Pottman, eds., *Eurographics Symposium on Geometry Processing 2005*, Vienna, Austria, 2005, pp. 149–158. Eurographics.

[73] Y. Ohtake, A. Belyaev, and H.-P. Seidel. An integrating approach to meshing scattered point data. In *ACM Symposium on Solid and Physical Modeling (SPM 2005)*, Cambridge, MA, USA, June 2005, pp. 61–69. ACM.

[74] S. Popov, J. Günther, H.-P. Seidel, and P. Slusallek. Experiences with streaming construction of SAH kd-trees. In I. Wald and S. G. Parker, eds., *Proceedings of the 2006 IEEE Symposium on Interactive Ray Tracing*, Salt Lake City, USA, September 2006, pp. 89–94. IEEE. Best Paper Award.

[75] A. Román and H. P. A. Lensch. Automatic multiperspective images. In T. Akenine-Möller and W. Heidrich, eds., *Rendering Techniques 2006: Eurographics Symposium on Rendering*, Nicosia, Cyprus, June 2006, pp. 161–171. Eurographics Association.

[76] B. Rosenhahn, T. Brox, D. Cremers, and H.-P. Seidel. A comparison of shape matching methods for contour based pose estimation. In R. Reulke, U. Eckhardt, B. Flach, U. Knauer, and K. Polthier, eds., *Combinatorial image analysis, 11th International Workshop, IWCIA 2006*, Berlin, Germany, 2006, *LNCS 4040*, pp. 263–276. Springer.

[77] B. Rosenhahn, T. Brox, and J. Weickert. Three-dimensional shape knowledge for joint image segmentation and pose estimation. In W. Kropatsch, R. Sablatnig, and A. Hanbury, eds., *Pattern recognition, 27th DAGM Symposium*, Vienna, Austria, September 2005, *LNCS 3663*, pp. 109–116. Springer.

[78] B. Rosenhahn, L. He, and R. Klette. Automatic human model generation. In A. Gagalowicz and W. Philips, eds., *Computer analysis of images and patterns, 11th International Conference, CAIP 2005, Versailles, France, September 5-8, 2005 Computer Analysis of Images and Patterns 11th Int. Conference*, Versailles, France, September 2005, *LNCS 3691*, pp. 41–48. Springer.

[79] B. Rosenhahn, H. Ho, and R. Klette. Texture driven pose estimation. In M. Sarfraz, Y. Wand, and E. Banissi, eds., *Computer Graphics, Imaging and Visualization, New Trends (CGIV 05)*, Beijing, China, 2005, pp. 271–277. IEEE.

[80] B. Rosenhahn, U. Kersting, K. Powell, and H.-P. Seidel. Cloth x-ray: MoCap of people wearing textiles. In K. Franke, K. R. Müller, B. Nickolay, and R. Schäfer, eds., *Pattern Recognition, 28th DAGM Symposium*, Berlin, Germany, September 2006, *LNCS 4174*, pp. 495–504. Springer.

[81] B. Rosenhahn, U. Kersting, A. Smith, J. Gurney, T. Brox, and R. Klette. A system for marker-less human motion estimation. In W. Kropatsch, R. Sablatnig, and A. Hanbury, eds., *Pattern recognition, 27th DAGM Symposium*, Vienna, Austria, September 2005, *LNCS 3663*, pp. 230–237. Springer.

[82] W. Saleem, D. Wang, A. Belyaev, and H.-P. Seidel. Statistical learning for shape applications. In *1st International Symposium on Shapes and Semantics*, Matsushima, JAPAN, June 2006, pp. 53–60. CNR.

[83] N. Sauber, H. Theisel, and H.-P. Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. In E. Gröller, A. Pang, C. T. Silva, J. Stasko, and J. van Wijk, eds., *IEEE Visualization Conference 2006*, Baltimore, USA, November 2006, *IEEE Visualization*, vol. 12, pp. 917–924. IEEE.

[84] O. Schall, A. Belyaev, and H.-P. Seidel. Robust filtering of noisy scattered point data. In M. Pauly and M. Zwicker, eds., *IEEE/Eurographics Symposium on Point-Based Graphics*, Stony Brook, New York, USA, 2005, pp. 71–77. Eurographics.

[85] O. Schall, A. Belyaev, and H.-P. Seidel. Adaptive fourier-based surface reconstruction. In M.-S. Kim and K. Shimada, eds., *Geometric modeling and processing - GMP 2006, 4th International Conference*, Pittsburgh, PA, USA, 2006, *LNCS 4077*, pp. 34–44. Springer.

[86] O. Schall, A. Belyaev, and H.-P. Seidel. Feature-preserving denoising of time-varying range data. In H. Pfister, ed., *SIGGRAPH 2006 Sketches and Applications*, Boston, Massachusetts, USA, 2006, p. 56. ACM.

[87] O. Schall, A. Belyaev, and H.-P. Seidel. Feature-preserving non-local denoising of static and time-varying range data. In *ACM Symposium on Solid and Physical Modeling*, Beijing, China, 2007. ACM. to appear.

[88] O. Schall and M. Samozino. Surface from scattered points: A brief survey of recent developments. In B. Falcidieno and N. Magnenat-Thalmann, eds., *1st International Workshop towards Semantic Virtual Environments*, Villars, Switzerland, 2005, pp. 138–147. MIRALab.

[89] G. Schlosser, J. Hesser, F. Zeilfelder, C. Rössl, R. Männer, G. Nürnberger, and H.-P. Seidel. Fast visualization by shear-warp on quadratic super-spline models using wavelet data decompositions. In C. T. Silva, E. Gröller, and H. Rushmeier, eds., *16th IEEE Visualization Conference (VIS 2005)*, Minneapolis, MN, USA, 2005, pp. 351–358. IEEE.

[90] B. Schölkopf, V. Blanz, and F. Steinke. Object correspondence as a machine learning problem. In *International Conference on Machine Learning ICML*, Bonn, Germany, 2005, pp. 776–783. ACM Press.

[91] B. Schölkopf, F. Steinke, and V. Blanz. Object correspondence as a machine learning problem. In *International Conference on Machine Learning ICML 2005*, Bonn, Germany, 2007.

[92] T. Schultz, B. Burgeth, and J. Weickert. Flexible segmentation and smoothing of DT-MRI fields through a customizable structure tensor. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, and T. Malzbender, eds., *Advances in Visual Computing, second International Symposium, ISVC 2006, Part I*, Lake Tahoe, NV, USA, October 2006, *LNCS 4291*, pp. 455–464. Springer.

[93] T. Schultz, H. Theisel, and H.-P. Seidel. Segmentation of DT-MRI anisotropy isosurfaces. In K. Museth, T. Möller, and A. Ynnerman, eds., *Proc. Eurographics / IEEE-VGTC Symposium on Visualization (EuroVis '07)*, Norrköping, Sweden, May 2007. Eurographics. Accepted for publication.

[94] K. Shi, H. Theisel, T. Weinkauf, H. Hauser, H.-C. Hege, and H.-P. Seidel. Path line oriented topology for periodic 2D time-dependent vector fields. In B. Sousa Santos, T. Ertl, K. I. Joy, D. W. Fellner, T. Möller, and S. N. Spencer, eds., *EUROVIS 2006, Eurographics / IEEE VGTC Symposium on Visualization*, Lisbon, Portugal, 2006, pp. 139–146. Eurographics.

[95] K. Smith, G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Beyond tone mapping: Enhanced depiction of tone mapped HDR images. In L. Szirmay-Kalos and E. Gröller, eds., *The European Association for Computer Graphics 27th Annual Conference, EUROGRAPHICS 2006*, Vienna, Austria, September 2006, *Computer Graphics Forum*, vol. 25, pp. 427–438. Blackwell.

[96] K. Smith, Y. Liu, and A. Klein. Animosaics. In K. Anjyo and P. Faloutsos, eds., *Computer Animation 2005, ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, Los Angeles, California, 2005, pp. 201–208. ACM.

[97] M. Smyk, S. Kinuwaki, R. Durikovic, and K. Myszkowski. Temporally coherent irradiance caching for high quality animation rendering. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, 2005, *Computer Graphics Forum*, vol. 24, pp. 401–412. Blackwell.

[98] G. Sommer, B. Rosenhahn, and C. Perwass. The twist representation of free-form objects. In R. Klette, R. Kozera, L. Noakes, and J. Weickert, eds., *Geometric Properties from Incomplete Data*, Dagstuhl, Wadern, Germany, 2006, *Computational Imaging and Vision*, vol. 31, pp. 3–22. Springer.

[99] F. Steinke, B. Schölkopf, and V. Blanz. Support vector machines for 3d shape processing. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, 2005, *Computer Graphics Forum*, vol. 24, pp. 285–294. Blackwell.

[100] C. Stoll, S. Gumhold, and H.-P. Seidel. Visualization with stylized line primitives. In C. T. Silva, E. Gröller, and H. Rushmeier, eds., *IEEE Visualization 2005 (VIS 2005)*, Minneapolis, USA, 2005, pp. 695–702. IEEE.

[101] C. Stoll, S. Gumhold, and H.-P. Seidel. Incremental raycasting of piecewise quadratic surfaces on the GPU. In I. Wald and S. G. Parke, eds., *IEEE Symposium on Interactive Raytracing 2006 Proceedings*, Salt Lake City, USA, September 2006, IEEE Symposium on Interactive Raytracing Proceedings, pp. 141–150. IEEE.

[102] C. Stoll, Z. Karni, C. Rössl, H. Yamauchi, and H.-P. Seidel. Template deformation for point cloud fitting. In M. Botsch and B. Chen, eds., *Symposium on Point-Based Graphics*, Boston, USA, 2006, pp. 27–35. Eurographics.

[103] C. Stoll, Z. Karni, and H.-P. Seidel. Geodesics guided constrained texture deformation. In *The 14th Pacific Conference on Computer Graphics and Applications Proceedings*, Taipei, Taiwan, October 2006, *Pacific Conference on Computer Graphics and Applications Proceedings*, vol. 14, pp. 144–152. National Taiwan University Press.

[104] C. Stoll, H.-P. Seidel, and M. Alexa. Bsp shapes. In *2006 International Conference on Shape Modeling and Applications (SMI 2006)*, Matsushima, Japan, June 2006, Proceedings of the IEEE International Conference on Shape Modeling and Applications, pp. 42–47. IEEE.

[105] H. Theisel, J. Sahner, T. Weinkauf, H.-C. Hege, and H.-P. Seidel. Extraction of parallel vector surfaces in 3d time-dependent fields and applications to vortex core line tracking. In C. T. Silva, E. Gröller, and H. Rushmeier, eds., *IEEE Visualization 2005 (VIS 2005)*, Minneapolis, USA, 2005, pp. 631–638. IEEE.

[106] C. Theobalt, M. Magnor, and H.-P. Seidel. 3D image analysis and synthesis at MPI informatik. In D. W. Fellner, ed., *Vision, Video and Graphics 2005 (VVG'05)*, Edinburgh, UK, July 2005, pp. 85–91. Eurographics.

[107] I. Wald, A. Dietrich, C. Benthin, A. Efremov, T. Dahmen, J. Günther, V. Havran, H.-P. Seidel, and P. Slusallek. A ray tracing based framework for high-quality virtual reality in industrial design applications. In I. Wald and S. G. Parker, eds., *Proceedings of the 2006 IEEE Symposium on Interactive Ray Tracing*, Salt Lake City, USA, September 2006, pp. 177–185. IEEE.

[108] I. Wald and H.-P. Seidel. Interactive ray tracing of point-based models. In M. Pauly and M. Zwicker, eds., *Symposium on Point-Based Graphics*, Stony Brook, USA, June 2005, pp. 9–16. Eurographics Association.

[109] T. Weinkauf, J. Sahner, H. Theisel, H.-C. Hege, and H.-P. Seidel. A unified feature extraction architecture. In *Active Flow Control*, Berlin, Germany, 2006. Collaborative Research Center 557.

[110] T. Weinkauf, H. Theisel, H.-C. Hege, and H.-P. Seidel. Topological structures in two-parameter-dependent 2D vector fields. In L. Szirmay-Kalos and E. Gröller, eds., *EUROGRAPHICS 2006 (EG'06)*, Vienna, Austria, 2006, *Computer Graphics Forum*, vol. 25, pp. 607–616. Blackwell.

[111] T. Weinkauf, H. Theisel, K. Shi, H.-C. Hege, and H.-P. Seidel. Extracting higher order critical points and topological simplification of 3d vector fields. In C. T. Silva, E. Gröller, and H. Rushmeier, eds., *IEEE Visualization 2005 (VIS 2005)*, Minneapolis, USA, 2005, pp. 559–566. IEEE.

[112] H. Yamauchi, S. Lee, Y. Lee, Y. Ohtake, A. Belyaev, and H.-P. Seidel. Feature sensitive mesh segmentation with mean shift. In M. Spagnuolo, A. Pasko, and A. Belyaev, eds., *Shape Modeling International 2005 (SMI 2005)*, Cambridge, MA, USA, June 2005, pp. 236–243. IEEE.

[113] H. Yamauchi, W. Saleem, S. Yoshizawa, Z. Karni, A. Belyaev, and H.-P. Seidel. Towards stable and salient multi-view representation of 3d shapes. In M. Spagnuolo, A. Belyaev, and H. Suzuki, eds., *IEEE International Conference on Shape Modeling and Applications 2006 (SMI 2006)*, Matsushima, JAPAN, June 2006, pp. 265–270. IEEE.

[114] M. Yoon, Y. Lee, S. Lee, I. Ivrissimtzis, and H.-P. Seidel. Ensembles for normal and surface reconstructions. In M.-S. Kim and K. Shimada, eds., *Geometric Modeling and Processing - GMP 2006, 4th International Conference,*, Pittsburgh, PA, USA, 2006, *LNCS 4077*, pp. 17–33. Springer.

[115] A. Yoshida, V. Blanz, K. Myszkowski, and H.-P. Seidel. Perceptual evaluation of tone mapping operators with real-world scenes. In *Human Vision and Electronic Imaging X, IS&T/SPIE's 17th Annual Symposium on Electronic Imaging (2005)*, San Jose, USA, January 2005, *SPIE Proceedings Series*, vol. 5666, pp. 192–203. SPIE.

[116] A. Yoshida, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Analysis of reproducing real-world appearance on displays of varying dynamic range. In *EUROGRAPHICS 2006 (EG'06)*, Vienna, Austria, September 2006, *Computer Graphics Forum*, vol. 25, pp. 415–426. Blackwell.

[117] S. Yoshizawa, A. Belyaev, and H.-P. Seidel. Fast and robust detection of crest lines on meshes. In L. Kobbelt and V. Shapiro, eds., *Proceedings of the Ninth ACM Symposium on Solid and Physical Modeling 2005*, Cambridge, Massachusetts, USA, 2005, pp. 227–232. ACM. Technical Sketch.

[118] S. Yoshizawa, A. Belyaev, and H.-P. Seidel. Smoothing by example: Mesh denoising by averaging with similarity-based weights. In M. Spagnuolo, A. Belyaev, and H. Suzuki, eds., *IEEE International Conference on Shape Modeling and Applications 2006 (SMI 2006)*, Matsushima, JAPAN, June 2006, pp. 38–44. IEEE.

[119] R. Zayer, C. Rössl, Z. Karni, and H.-P. Seidel. Harmonic guidance for surface deformation. In M. Alexa and J. Marks, eds., *The European Association for Computer Graphics 26th Annual Conference, EUROGRAPHICS 2005*, Dublin, Ireland, 2005, *Computer Graphics Forum*, vol. 24, pp. 601–609. Blackwell.

[120] R. Zayer, C. Rössl, and H.-P. Seidel. Discrete tensorial quasi-harmonic maps. In M. Spagnuolo, A. Pasko, and A. Belyaev, eds., *Shape Modeling International 2005 (SMI 2005)*, Cambridge, MA, U.S.A, 2005, pp. 276–285. IEEE.

[121] R. Zayer, C. Rössl, and H.-P. Seidel. Setting the boundary free: A composite approach to surface parameterization. In M. Desbrun and H. Pottmann, eds., *Symposium on Geometry Processing*, Vienna, Austria, 2005, pp. 91–100. Eurographics/ACM.

[122] R. Zayer, C. Rössl, and H.-P. Seidel. Curvilinear spherical parameterization. In *Shape Modeling International (SMI)*, Matsushima, Japan, 2006, pp. 57–64. IEEE.

[123] Y. Zhao, R. J. Valkenburg, R. Klette, and B. Rosenhahn. Target calibration and tracking using conformal geometric algebra. In L.-W. Chang and W.-N. Lie, eds., *Advances in Image and Video Technology, First Pacific Rim Symposium, PSIVT 2006*, Hsinchu, Taiwan, December 2006, *LNCS 4319*, pp. 74–83. Springer.

[124] G. Ziegler, R. Dimitrov, C. Theobalt, and H.-P. Seidel. Real-time quadtree analysis using histopyramids. In B. E. Rogowitz and T. N. Pappas, eds., *IS&T and SPIE Conference on Electronic Imaging*, San Jose, USA, January 2007, Proceedings of SPIE-IS&T Electronic Imaging. SPIE and IS&T.

[125] G. Ziegler, M. Magnor, and H.-P. Seidel. Geocast: Unifying depth video with camera meta-data. In C. Weigel, P. Schübel, and D. F. Harezlak, eds., *2nd Workshop on Immersive Communication and Broadcast Systems*, Berlin, Germany, October 2005, Conference CD, pp. 0–4. Heinrich-Hertz-Institute.

[126] G. Ziegler, C. Theobalt, and H.-P. Seidel. On-the-fly point clouds through histogram pyramids. In L. Kobbelt, T. Kuhlen, T. Aach, and R. Westermann, eds., *11th International Fall Workshop on Vision, Modeling and Visualization 2006 (VMV2006)*, Aachen, Germany, 2006, pp. 137–144. Aka.

16 The Databases and Information Systems Group (D5)

16.1 Personnel

Director

Prof. Dr.-Ing. Gerhard Weikum

Researchers

Srikanta Bedathur Jagannath, PhD (July 2005–)
Mouna Kacimi, PhD (December 2006–)
Dr. Thomas Neumann
Maya Ramanath, PhD (July 2005–)
Dr.-Ing. Ralf Schenkel
Dr.-Ing. Stefan Siersdorfer (–December 2006, now University of Sheffield)
Mauro Sozio (January 2007–)
Christos Tryfonopoulos, PhD (September 2006–)

PhD Students

Ralitsa Angelova
Matthias Bender
Klaus Berberich
Andreas Broschart (January 2006–)
Tom Crecelius (January 2006–)
Gerard de Melo (January 2007–)
Jens Graupmann (–August 2006, now IMG GmbH, Munich)
Gjergji Kasneci (March 2006–)
Georgiana Ifrim
Julia Luxenburger
Sebastian Michel
Hanglin Pan
Josiane Xavier Parreira
German Shegalov (–July 2005, now Oracle Inc., Portland)
Sergej Sizov (–December 2005, now Koblenz University)
Fabian Suchanek (August 2005–)
Martin Theobald (–September 2006, now Stanford University)
Christian Zimmer

Secretaries
Adriana Davidescu
Petra Schaaf

16.2 Visitors

In the time period from June 2005 to March 2007, the following researchers visited our group:

Peter Triantafillou	15.09.04-15.07.05 11.12.05-17.12.05	University of Patras
Debora Donato	15.05.05-16.08.05	University of Rome "La Sapienza"
Nikos Ntarmos	16.05.05-22.05.05	University of Patras
P. Sreenivasa Kumar	05.11.05.-11.11.05 29.05.06-07.07.06	IIT Madras
K. Hima Prasad	05.11.05-11.11.05	IIT Madras
Andre Seifert	09.11.05-10.11.05	Uni Konstanz
David Novak	13.11.05-03.12.05	University of Brno
Christos Tryfonopoulos	17.11.05-19.12.05 19.02.06-24.02.06	Technical University of Crete
Nikos Ntarnos	11.12.05-17.12.05	University of Patras
Andreas Henrich	19.12.05	University of Bamberg
Tobias Scheffer	21.12.05-23.12.05	Humboldt-University Berlin
Seung-Won Hwang	02.01.06-27.02.06 02.01.07-28.02.07	POSTECH, Korea
Arun Mukhija	11.01.06	Zurich University
David Lomet	16.03.06-19.03.06	Microsoft Research
Alessandro Linari	25.03.06-08.10.06	University of Bologna
Thomas Hofmann	19.04.06	TU Darmstadt
Julia Stoyanovich	05.06.06-31.08.06	Columbia University
Norbert Fuhr	15.05.06-16.05.06	University of Duisburg-Essen
Michalis Vazirgiannis	15.05.06-18.05.06 03.07.06-14.07.06	AUEB, Athens
Georgia Koutrika	05.06.06-07.06.06	University of Athens
Reinhard Walker	06.06.06	juris GmbH, Saarbrücken
Felix Weigel	06.06.06-08.06.06	University of Munich
Alkis Simitsis	25.06.06-27.06.06	National Technical University of Athens
Mouna Kacimi	04.09.06-05.09.06	Université de Bourgogne
Mauro Sozio	16.10.06-18.10.06	University of Rome "La Sapienza"
Tobias Scheffer	20.10.06	Humboldt-University, Berlin
Thomas Roelleke	23.10.06-03.11.06	Queen Mary University of London

Volker Markl	10.11.06-11.11.06	IBM Almaden Research Center
Dean Jacobs	10.11.06-11.11.06	TU Munich
Marco Patella	13.11.06-15.11.06	University of Bologna
Alessandro Linari	13.11.06-18.11.06	University of Bologna
Sven Helmer	24.11.06	Birkbeck College, London
Etzard Stolte	29.11.06-30.11.06	Unilever R&D, The Netherlands
Donald Kossmann	30.11.06	ETH Zurich
Christian Duda	30.11.06	ETH Zurich
Yosi Mass	20.12.06-21.12.06	IBM Haifa
Bernhard Seeger	26.03.-04.05.07	University of Marburg

16.3 Group Organization

The group's research falls into four major areas:

1. semantic search and semistructured data management,
2. peer-to-peer search and information management,
3. query processing, and
4. Web and text mining.

Each area is coordinated by an experienced researcher or senior postdoc, and involves 4 to 8 researchers with overlaps among areas. The director and the coordinators have weekly meetings.

The first two areas involve the development of major prototype systems, TopX and NAGA for semantic search and Minerva for peer-to-peer search. These serve for experimentation and integration, and are also catalysts for new research ideas. TopX and Minerva are available as open-source software. There is fruitful interaction among the four areas. For example, our results on query processing are used for efficient search, and new methods for Web and text mining contribute to the expressiveness and effectiveness of semantic search.

Notwithstanding the organization into four areas, the department has a collaborative rather than hierarchical structure. All group members report to the director. Each student and researcher writes a short progress report every six months, which is discussed with the director. All graduate students meet regularly with the director and also interact with other senior scientists at the institute or the Saarland University, who could potentially serve as supervisors and readers of the doctoral theses. The group runs a weekly research seminar with informal presentations and discussion, and has also been running a biweekly reading group, with focus on link analysis (for Web graph and social networks) in the timeframe 2005–2006 and focus on applied machine learning (for classification, tagging, grouping, ranking, and specific NLP tasks) in 2006–2007. The latter involves interactions with the institute's Machine Learning group (RG2). In addition, members of D5 participate in the semi-annual Oberseminar, which consists of presentations on Diploma and Master thesis projects, and various activities within the International Max Planck Research School (IMPRS) and the Graduate Studies Program (Graduiertenkolleg) at the Saarland University.

16.4 Semantic Search and Semistructured Data Management

Coordinator: Ralf Schenkel

16.4.1 XML IR

Investigators: Ralf Schenkel, Martin Theobald, and Gerhard Weikum

Query languages for XML such as XPath or XQuery support Boolean retrieval: a query result is a (possibly restructured) subset of XML elements or entire documents that satisfy the search conditions of the query. This search paradigm works for highly schematic XML data collections such as electronic catalogs. However, non-schematic XML data that comes from many different sources and inevitably exhibits heterogeneous structures and annotations (i.e., XML tags) cannot be adequately searched using such database query languages, as queries often either return too many or too few results. 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. These considerations also hold for applications that deal with structured data but face uncertainty in the data, for example, because the data is probabilistic or the data is derived from text, Web, and other sources by automatic information extraction methods or by heuristically reconciling many different databases that vary in their schemas and representations of entities. For example, we have converted the IMDB movie database, a structured dataset, and the Wikipedia encyclopedia, a hyperlinked text corpus, into XML as test cases for ranked retrieval of XML data.

Example queries are (in the W3C XPath Full-Text syntax)

```
/movie[.//location ftcontains "university"]
      [.//plot ftcontains "mathematics"]//casting
```

or

```
/article[.//topic ftcontains "travel"]
        //section[.//caption ftcontains "church"].
```

In both cases, ranking the search results becomes important if there are no results that perfectly match all query conditions, or if there are too many perfect matches of different quality, or if the user is willing to relax some of the content conditions, e.g., finding "college" or "Princeton" instead of "university", or some of the structural constraints, e.g., finding articles that contain only "category" tags instead of "topic" tags and not having any "section" elements but matching the other conditions.

TopX

Our XML Search Engine TopX [5, 7] provides powerful search and ranking functionality for this class of applications, addressing recent trends towards integrating database systems (DB) and information retrieval (IR) [1]. From a DB viewpoint, TopX provides an efficient algorithmic basis for top-k query processing over multidimensional datasets, ranging from structured data such as product catalogs (e.g., bookstores, real estate, movies, etc.) to unstructured text documents (with keywords or stemmed terms defining the feature space) and semistructured XML data in between. From an IR viewpoint, TopX provides ranked retrieval based on a relevance scoring function, with support for flexible combinations of mandatory and optional conditions as well as text predicates such as phrases, negations,

etc. TopX combines these two aspects into a unified framework and software system, with emphasis on XML ranked retrieval.

The research centered around TopX makes the following major contributions:

- comprehensive support for the ranked retrieval functionality of XPath Full-Text, including a probabilistic-IR scoring model for full-text content conditions and tag-term combinations, path conditions for all XPath axes as exact or relaxable constraints, and ontology-based relaxation of terms and tag names as similarity conditions for ranked retrieval,
- efficient and scalable techniques for content-and-structure indexing and query processing, with demonstrated good performance on large-scale benchmarks,
- probabilistic models for approximate top-k query processing that predict scores in sequential index scans and can thus accelerate queries by earlier termination and lower memory consumption [8],
- judicious scheduling of sequential and random index accesses for further run-time improvements [3], specifically designed for handling XML data, and
- efficient support for integrating ontologies and thesauri, by incremental merging of index lists for on-demand, self-throttling query expansion[4].

Figure 16.1 depicts the TopX main components. TopX comes with Web and file crawlers for indexing text and XML data in a relational schema on top of a database engine used as a storage platform. TopX currently uses Oracle 10g but could easily be ported to other DBMSs or a customized index using inverted files. The TopX core query processor provides the algorithms for efficient top-k query evaluation with early termination, with the option of plugging in specialized components for probabilistic candidate pruning, cost-oriented index access scheduling for further query acceleration, and dynamic query expansion for IR-style relaxed search. The query processor operates in a multi-threaded way with asynchronous disk I/O for best possible performance of sequential scans.

TopX provides both an interactive Web frontend for human users and a Web Service API to be used by other applications. A demo of the system is available at `http://infao5501.ag5.mpi-sb.mpg.de:8080/topx`. TopX is completely implemented in Java as a servlet and class library running in a Tomcat server, with a code base of approximately 37,500 lines of code in 198 classes. It is available as open-source package from `http://topx.sourceforge.net`.

TopX has been extensively evaluated with the two major benchmark series in IR, namely the largest available collections of the Text Retrieval Conference (TREC) on text IR [2] and the Initiative for the Evaluation of XML Retrieval (INEX) on XML IR. For INEX, TopX performed especially well for content-and-structure queries, ranking among the top-5 of 25 in the 2005 benchmark, with a peak position 1 for two of the five official evaluation methods [6]. TopX has been the official host for the INEX 2006 topic development phase, and its Web Service interface is used by the INEX 2006 Interactive Track. During the topic development phase on an XML version of Wikipedia, more than 10,000 queries from roughly 70 different participants were processed.

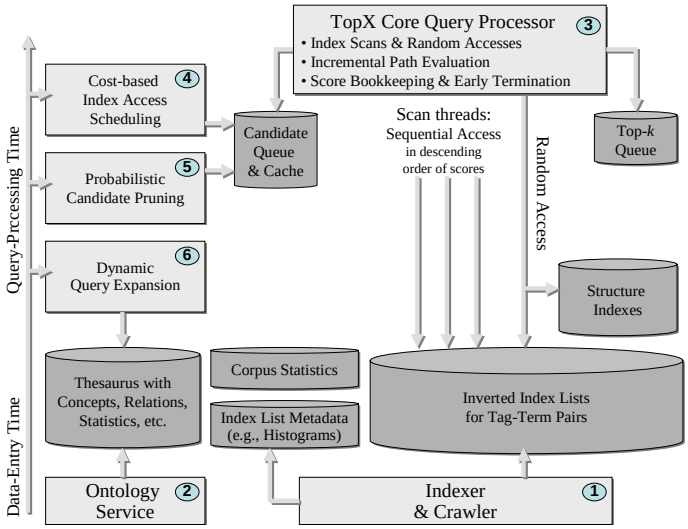


Figure 16.1: TopX Architecture

References

- [1] S. Amer-Yahia, P. Case, T. Roelleke, J. Shanmugasundaram, and G. Weikum. Report on the DB/IR panel at SIGMOD 2005. *SIGMOD Record*, 34(4):71–74, December 2005.

- [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] 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.

- [5] 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.

- [6] M. Theobald, R. Schenkel, and G. Weikum. TopX & XXL at INEX 2005 (ad-hoc track). In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, eds., *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, Dagstuhl Castle, Germany, 2006, *LNCS 3977*, pp. 282–295. Springer.

- [7] 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. ACM.

[8] M. Theobald, G. Weikum, and R. Schenkel. Top-k query evaluation with probabilistic guarantees. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, eds., *Proceedings 2004 VLDB Conference, The 30th International Conference on Very Large Databases (VLDB)*, Toronto, Canada, 2004, pp. 648–659. Morgan Kaufmann. Acceptance ratio 1:6.

16.4.2 Relevance Feedback for XML IR

Investigators: Hanglin Pan, Ralf Schenkel, and Martin Theobald

Structural Relevance Feedback. XML search engines employ the ranked retrieval paradigm for producing relevance-ordered result lists rather than merely using XPath or XQuery for Boolean retrieval. However, even though there are extensions to these structured query languages that allow ranking based on textual conditions, the predominant paradigm for user-side query formulation has been, like in text retrieval, simple keyword queries. On one hand, this is due to the complexity of such structured query languages, which makes them infeasible as query language for end users. On the other hand, the schema of the data collection may not be known to the user or may be highly heterogeneous. As a consequence of their simplicity, keyword queries cannot exploit the rich annotations available in XML, so the results of an initial query are often not very satisfying.

Relevance Feedback is an important way to enhance retrieval quality by integrating relevance information provided by a user. In XML retrieval, existing feedback engines usually generate an expanded keyword query from the content of elements marked as relevant or non-relevant. This approach that is inspired by text-based IR completely ignores the semistructured nature of XML. We have made the important step from content-based to structural feedback, by extending the well-established feedback approaches by Rocchio [3] and Robertson and Sparck-Jones [2] to expand a keyword-based query into a possibly complex content-and-structure query that specifies new constraints on the structure of results, in addition to "standard" content-based query expansion. The resulting expanded query has weighted structural and content constraints and can be fed into a full-fledged XML search engine like our own TopX engine (see Section 16.4.1).

As an example, consider the keyword query `moldovan semantic networks`. Without additional knowledge, it is unclear that the term "moldovan" actually refers to the author of a paper about semantic networks. Additionally, it is very unlikely that the author name and the terms "semantic network" occur in the same element, as author names are usually mentioned in different places than the content of articles. A query with constraints on both content and structure would probably yield a lot more relevant results, but it is impossible to formulate a query like the following without knowledge of the underlying schema:

```
//sec[about(.//author, "moldovan") and about(., "semantic networks")]
```

Our relevance feedback framework [5, 7] automatically constructs a content-and-structure query from a keyword-based query, exploiting relevance feedback by a user. Besides the content of relevant elements, our framework considers also tags and content of other elements related to relevant elements, such as ancestors and descendants, but also descendants of ancestors (to include, for example, references in a paper).

We successfully evaluated our feedback framework within the Relevance Feedback task of the established INEX benchmark. Both 2005 and 2006, our approach managed to yield

better relative and absolute improvements in query performance than the competitors, with a peak relative improvement of more than 25% [6, 10].

Our current work in this area focuses on exploiting advanced user feedback that goes beyond simple per-result feedback. Instead of supplying a single relevance value for a complete result element, the user should be allowed to 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 [4] 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 [8]. We then use the Transferable Belief Model [9] to compute, for each refinement candidate, a probability that it can identify relevant results.

Evaluation of Relevance Feedback Algorithms. The evaluation of feedback runs for XML IR is a problem that has not yet been solved in a satisfying way. The INEX Relevance Feedback task [1], the most authoritative community effort in this field, reuses collections, topics, result assessments and performance metrics from the INEX benchmark for XML retrieval. To evaluate the quality of a single feedback algorithm running on top of an XML search engine, a set of results (called *run*) computed by the search engine without feedback serves as a baseline. Then, for each of the top-k results for each topic, relevance information is extracted from the assessments and fed as input to the relevance feedback algorithm, which generates modified queries based on this "automatic" feedback. However, people have agreed that simply comparing the results of this run created from the modified queries to the baseline run is unfair as the new run has information about relevant elements and hence is biased. The most commonly used means to alleviate these effects is *freezing* the results used for feedback (and therefore with know relevance), thus assessing only the effect of reranking the results with unknown relevance.

However, freezing has some problems. First, as it does not change the first k results, any effect of the feedback algorithms can only be measured starting from the $k+1$'th element, reducing the absolute differences (and possibly also the significance). Second, this technique cannot be applied if not the top-k results are used for feedback, but other results (like only results from different documents, or results that are not necessarily part of the baseline run). We therefore propose to extend the so-called *residual collection* approach from text retrieval to XML. Here, results used for feedback are virtually removed from the collection, and the performance of both the baseline and the feedback run are measured against this residual collection only. However, as the elements in a document are not necessarily independent (if a section element is relevant, the corresponding article element will most likely be relevant, too), we need to remove related elements, too. We introduced and compared several variants of the residual collection technique in [6, 10], using a set of software tools that we developed and that was also used to compute the official results of the INEX 2006 Relevance Feedback Task.

While these methods enable us to compare the effects of different relevance feedback

algorithms that are run on top of a single search engine, it is still an open problem how to compare different feedback algorithms in a benchmark setting, i.e., when different feedback algorithms run on top of different retrieval engines. Here, the performance of a feedback algorithm is influenced by the performance of the underlying search engine. Search engines with a good performance will usually lead to an advantage for feedback algorithms running on top of them, as they will have more relevant results in the top-k of the baseline runs. A simple solution would be to use the same search engine for each feedback algorithm, but this is not always possible as some algorithms are tightly coupled with a certain engine (for example algorithms that modify internal weights or other parameters of the search engine). Our current work in this area focuses on providing each algorithm with the same set of results for feedback.

References

[1] C. J. Crouch. Relevance feedback at inex 2005. *SIGIR Forum*, 40(1):58–59, 2006.

[2] S. Robertson and K. Sparck-Jones. Relevance weighting of search terms. *Journal of the American Society of Information Science*, 27:129–146, May–June 1976.

[3] J. Rocchio Jr. Relevance feedback in information retrieval. In G. Salton, ed., *The SMART Retrieval System: Experiments in Automatic Document Processing*, ch. 14, pp. 313–323. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1971.

[4] 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.

- [5] R. Schenkel and M. Theobald. Feedback-driven structural query expansion for ranked retrieval of XML data. In Y. Ioannidis, M. H. Scholl, J. W. Schmidt, F. Matthes, M. Hatzopoulos, K. Boehm, A. Kemper, T. Grust, and C. Boehm, eds., *Advances in Database Technology - EDBT 2006: 10th International Conference on Extending Database Technology*, Munich, Germany, 2006, *LNCS 3896*, pp. 331–348. Springer. Acceptance ratio 1:6.

- [6] R. Schenkel and M. Theobald. Relevance feedback for structural query expansion. In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, eds., *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, Dagstuhl Castle, Germany, 2006, *LNCS 3977*, pp. 344–357. Springer.

- [7] 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.

[8] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.

[9] P. Shets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–243, 1994.

- [10] 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., *Preproceedings of the 5th International Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2006)*, Schloß Dagstuhl, Germany, 2006, pp. 140–149.

16.4.3 Proximity-Aware Ranking

Investigators: Andreas Broschart, Ralf Schenkel in collaboration with Seung-Won Hwang

Search engines make use of scoring functions in order to rank results to user queries. Hence, the quality of the scoring function is crucial to user satisfaction and success of the search engine. Content-based scoring functions such as Okapi BM25 [6] are based on the "bag of words" model, and thus refrain from considering *term proximity*, i.e., distances between query term occurrences in a document. Ignoring this information, however, frequently causes results that are unsatisfactory to users. As an example, suppose a user poses the query *surface area* of a *rectangular pyramid*. Scoring functions that do not consider proximity information of query term occurrences might return documents in which all of the query terms are individually important but appear in different paragraphs, e.g. treat "volume of a pyramid" in the first paragraph and "surface area of a rectangular prism" in the second. From a user's viewpoint, phrase queries are a way to avoid such bad results, but strict phrase matching would prevent the retrieval of many potentially good results. As an example, an occurrence of "surface area of a pyramid with rectangular base" would not be a hit, though would probably indicate a good result. Proximity-enhanced scoring functions that are based on positional information of query term occurrences are a mean to alleviate such effects. They provide the means to execute implicit soft phrases queries, such that the user does not have to specify them.

There are several proposals for proximity-enhanced scoring functions that loosen the rigidity of phrase queries by taking positional information of query term occurrences into account. They can be categorized into two classes: (1) linear combinations of a content score and a proximity score based on distance or size of a text window (e.g., [1, 4, 5]), and (2) integrated scoring models (e.g., [3, 7]). We have carried out extensive experimental studies with the TREC Terabyte and Robust queries to determine which of these proximity-aware scoring functions perform well. In our experiments, it turned out that the approaches suggested by Büttcher et al.[1] and Song et al.[7] yield statistically significant improvements in result quality on the TREC Terabyte corpus with the 100 topics from the TREC 2004 and TREC 2005 Terabyte AdHoc track compared to Okapi BM25, whereas the other approaches could not improve or even impaired result quality.

The existing approaches for proximity-aware retrieval focus on retrieval effectiveness but not on efficient query processing, so it is often unclear how they could be efficiently implemented in a real search engine. We intent to overcome this issue by means of top-k query processing, which aims at efficiently computing only the best k results. Here, the method of choice is the family of threshold algorithms (TA) [2] that perform index scans over precomputed index structures and aggregate scores on the fly in order to compute a lower bound for the current k-ranked result and an upper bound for all candidates that are not in the top-k. If the lower bound for the current k-ranked result, the threshold, is at least as high as the upper bounds of all candidates, the algorithm can terminate. Often, this is long before the index lists have been scanned completely. Besides these sequential accesses (SA) to indexes, carefully selected random accesses (RA) can make the algorithm terminate even earlier, but the cost for each random access c_{RA} is 50–50,000 times higher than the cost for each sequential access c_{SA}.

We have implemented and evaluated first top-k variants of Büttcher's scoring function [1]

within our TopX engine (see Section 16.4.1) that utilize precomputed proximity scores for term pairs in combination with the usual index lists for terms. In our experiments we lowered the query processing costs by up to two orders of magnitude using early pruning and combined index lists. These very promising results raise our hopes that better index structures can further improve the results.

Our future research in this area aims at deriving proximity-enhanced scoring functions for more sophisticated tasks, including (1) semi-structured retrieval based on (a) keyword queries for XML documents and (b) structural queries for XML documents, and (2) graph retrieval. We will additionally consider efficient algorithms for pruning the index size with controllable guarantees on retrieval quality.

References

[1] S. Büttcher et al. Term proximity scoring for ad-hoc retrieval on very large text collections. In *SIGIR*, 2006, pp. 621–622.

[2] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *J. Comput. Syst. Sci.*, 66(4):614–656, 2003.

[3] O. de Kretser and A. Moffat. Effective document presentation with a locality-based similarity heuristic. In *SIGIR*, 1999, pp. 113–120. ACM.

[4] C. Monz. Minimal span weighting retrieval for question answering. In *SIGIR Workshop on Information Retrieval for Question Answering*, 2004, pp. 23–30.

[5] Y. Rasolofo and J. Savoy. Term proximity scoring for keyword-based retrieval systems. In F. Sebastiani, ed., *ECIR*, 2003, *LNCS 2633*, pp. 207–218. Springer.

[6] S. E. Robertson et al. Okapi at TREC-3. In *TREC*, 1994, pp. 109–126.

[7] R. Song et al. Viewing term proximity from a different perspective. Technical Report MSR-TR-2005-69, Microsoft Research Asia, May 2005.

16.4.4 Graph-Based IR

Investigators: Jens Graupmann, Georgiana Ifrim, Gjergji Kasneci, Maya Ramanath, Ralf Schenkel, Fabian Suchanek, and Gerhard Weikum

SphereSearch

Today's web search engines are still following the paradigm of keyword based search. Although this is the best choice for large scale search engines, such as Google, in terms of throughput and scalability, it inherently limits their abilities to accomplish more meaningful query tasks. On the other hand, XML query engines (e.g., based on XQuery or XPath) have powerful query capabilities but at the same time their dedication to XML data with a global schema is their weakness, due to the fact that most web information is still stored in diverse formats and does not conform to common schemas.

The SphereSearch Engine [5, 3] provides unified ranked retrieval on these heterogeneous XML and Web data. To obtain a unified view on all documents, regardless of their original data format (HTML, PDF, etc), all documents are automatically converted into XML format considering their semantic structure. In contrast to semi-automatic wrapper approaches this

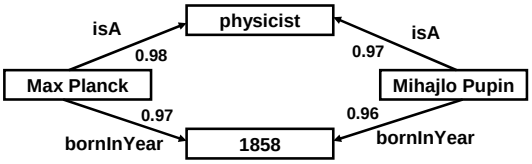
conversion is done fully automatically using heuristics and machine-learning techniques for tables and forms [2]. Additionally linguistic annotation tools like GATE [1] are integrated to annotate named entities like locations and persons, thus adding further semantics and structure.

We developed a query language that allows us to search within heterogeneous XML and Web data as well as combinations in a unified manner. The language is implemented in the SphereSearch Engine. Its design has been influenced by our earlier work on the XXL language for XML IR [8], on one hand, and the desire to handle also current Web data in HTML, on the other hand. But SphereSearch also deviates from and significantly extends prior work by interpreting all data as a graph structure rather than trees, with XPath-style search conditions across document/page boundaries and a scoring/ranking model that reflects the compactness of a matching subgraph.

The SphereSearch Engine is fully implemented in Java using Oracle10g as an underlying data manager. Besides a servlet-based user interface for query formulation and result presentation, it can also illustrate results of geographical queries in GoogleEarth, using additional geographic information from gazetteers [4]. We have carried out experiments on large-scale datasets like the well-established INEX benchmark for XML-IR, the open Internet encyclopedia Wikipedia consisting of more than 600,000, highly cross-linked lexicon entries in combination with structured data from IMDB, and the DBLP data converted to XML in combination with href links to researcher homepages and further Web pages about projects, courses, etc. Our experiments demonstrate both the system's efficiency and its expressiveness and search result quality.

NAGA

Current keyword-oriented search engines for the World Wide Web do not allow specifying the *semantics* of queries. As a concrete example, suppose we want to find out which other physicists were born in the same year as Max Planck. First, it is close to impossible to formulate this query in terms of keywords. Second, the 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 concepts corresponding to query terms or relations holding between the query terms. Thus, the preceding query could be expressed as:



We address the problem above with NAGA[1], a new semantic search engine, which builds on a large semantic knowledge base of binary relationships (facts) derived from the Web.

NAGA's data model is a graph, in which nodes are labeled with *entities* (e.g. `Max Planck`) and edges are labeled with relationships (e.g. `bornInYear`). Each edge, together with its end

[1] In the mythologies of Hinduism and Buddhism, NAGA is a huge snake. Here, it stands for the size and diversity of the Web.

nodes, represents a fact, e.g. `Max Planck bornInYear 1858` or `Max Planck type physicist`. Since these facts are derived from Web pages using possibly unreliable Information Extraction techniques, we attach a *certainty value* to each fact. We compute the certainty value $c(e) \in [0,1]$ for an edge e as:

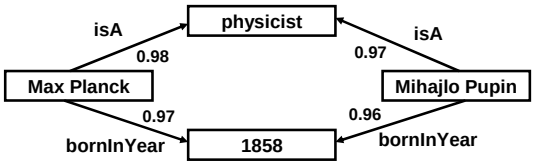
$$c(e) = \sum_{i=1}^{n_e} C(e, P_i) T(P_i)$$

where P_i denotes one of the n_e pages from which the fact corresponding to e was derived. The trust value $T(P_i)$ represents the authority of page P_i and can be computed by PageRank or similar algorithms. We assume $\sum_i T(P_i) = 1$. $C(e, P_i)$ is the confidence with which the fact corresponding to e was extracted from page P_i. Thus, the certainty value for e accumulates trust and extraction quality values across all pages in which the corresponding fact was found (evidence pages). For each fact we maintain all the URLs of its evidence pages.

NAGA's knowledge base is a projection of YAGO (see Section 16.7.3).

In the spirit of the preceding example query, NAGA provides a taxonomy of graph queries in which edges or nodes may be unlabeled or edges may be labeled with regular expressions over relation names.

The answer to a query is a subgraph of the knowledge graph that matches the query. A possible answer to our example query, as returned by NAGA is:



In order to compute the overall score of an answer, we depart from language models for information retrieval. Given query $q = q_1 q_2 ... q_n$ and a candidate graph $g = g_1 g_2 ... g_n$, where each q_i is of the form (x, y, z) and at least one of the components is not a variable (unlabeled node or edge), we estimate the conditional probability that g generates q as:

$$P(g|q) \sim P(q|g)P(g)$$

where $P(g)$ can reflect a prior belief that g is relevant to any query. $P(q|g)$ is the query likelihood given the graph g, which captures how well the graph fits the particular query q. In our setting we assume $P(g)$ to be uniform and thus we are interested in computing $P(q|g)$. We assume probabilistic independence between the query's fact templates, which results in

$$P(q|g) = \prod_{i=1}^{n} P(q_i|g).$$

Following the approach in [6] adapted to our setting, we write the likelihood of a query fact given an answer graph as a mixture of two distributions, $\tilde{P}(q_i|g)$ and $\tilde{P}(q_i)$ as follows:

$$P(q_i|g) \quad = \quad \alpha \cdot \tilde{P}(q_i|g) + (1-\alpha) \cdot \tilde{P}(q_i),\ 0 \leq \alpha \leq 1 \tag{16.1}$$

$\tilde{P}(q_i|g)$ is the probability of drawing q_i randomly from an answer graph, $\tilde{P}(q_i)$ is the probability of drawing q_i randomly from the total knowledge graph and α is either automatically

learned (when relevance information about answer graphs is available [6]) or set to an empirically calibrated global value.

The connection between this formulation and $tf \cdot idf$ style retrieval models is shown in [6]). By means of our scoring model, we want to capture *confidence*, *informativeness*, and *compactness*. The two former desiderata are captured in the following formula:

$$\begin{aligned} \tilde{P}(q_i|g) &= \beta \cdot P_{conf}(q_i|g) + (1-\beta) \cdot P_{info}(q_i|g), \\ 0 \leq \beta \leq 1 \end{aligned} \tag{16.2}$$

Note that confidence and informativeness are indeed independent criteria. For example, we can be very confident that Albert Einstein was both a physicist and a politician, but the former fact is more informative than the latter, because Einstein was a physicist to a larger extent than he was a politician.

The maximum likelihood estimator for $P_{conf}(q_i|g)$ is given by:

$$P_{conf}(q_i|g) = \prod_{f \in match(q_i,g)} P(f \text{ holds}) \tag{16.3}$$

where $P(f \text{ holds})$ can be approximated by $c(f)$. If q_i is labeled by a simple relation, then $match(q_i, g)$ contains just one fact and $P_{conf}(q_i|g)$ is the confidence of that fact. If q_i is labeled with `connect` or with a regular expression over relations, then $match(q_i, g)$ contains the sequence of facts that together match q_i. The probability of that sequence being true is the product of the confidences of the single facts, assuming that the confidences of the single facts are independent.

The *informativeness* of a query template q_i given the answer graph g depends on the informativeness of each matching fact in g:

$$P_{info}(q_i|g) = \prod_{f \in match(q_i,g)} P_{finfo}(f|q_i) \tag{16.4}$$

The informativeness of the fact f depends on the unbound arguments of the query template q_i.

Let $f = (x, r, y)$ be an observation drawn from the joint distribution of three random variables X, R and Y. X and Y take values from the set of knowledge graph nodes and R takes values from the set of edges (i.e. relations). Given a query of the form $q_i = (x', r', y')$, if $f = (x, r, y)$ is a match for q_i, we define the informativeness of f as follows:

$$P_{finfo}(f|q_i) = \begin{cases} P(x|r,y), & \text{if } x' \text{ unbound in } q_i \\ P(y|r,x), & \text{if } y' \text{ unbound in } q_i \\ P(r|x,y), & \text{if } r' \text{ unbound in } q_i \\ P(x,y|r), & \text{if } x', y' \text{ unbound in } q_i \\ P(x,r|y), & \text{if } x', r' \text{ unbound in } q_i \\ P(r,y|x), & \text{if } r', y' \text{ unbound in } q_i \\ P(x,r,y), & \text{else} \end{cases} \tag{16.5}$$

We show how to estimate these probabilities by the example of $P(x|r,y)$. $P(x|r,y)$ can be written as follows:

$$P(x|r,y) = \frac{P(x,r,y)}{P(r,y)} = \frac{P(x,r,y)}{\sum_{x'} P(x',r,y)} \tag{16.6}$$

We estimate $P(x,r,y)$ using the number of witness pages for the fact (x,r,y) [2]:

$$P(x,r,y) \approx \frac{|W(x,r,y)|}{\sum_{x',r',y'} |W(x',r',y')|} \tag{16.7}$$

The third desideratum, *compactness* of results, is implicitly captured by this model. This is because the likelihood of an answer is the product over the likelihoods of its component facts. Thus, it is inversely correlated to the path length.

We turn to estimating $\tilde{P}(q_i)$. $\tilde{P}(q_i)$ plays the role of giving different weights to different components (i.e. triples) in the query. This is similar in spirit to the idf-methodology for weighting different query terms in traditional language models. Query triples are assigned weights which are inversely correlated to their frequency in the knowledge graph. Note that since the knowledge graph is virtually free of redundancy – unlike a document-level corpus, reasoning about background models and idf-style aspects more subtle and difficult.

To assess the quality of results returned by NAGA, we manually evaluated the output of 55 benchmark queries from the TREC 2005 and 2006 Question Answering tracks with NAGA (where we manually translated the natural language questions to NAGA's query language), Google, and Yahoo! Answers [7]. It is clear that NAGA had an advantage over Google and Yahoo! Answers, because the questions are already translated into the graph query language. At the same time, Google and Yahoo! Answers have a massive advantage over NAGA, because they are commercially operated systems that can search the whole Web (Google) or a huge corpus of several million predefined questions (Yahoo! Answers). We found that NAGA performs significantly better than Google for the TREC queries, whereas Yahoo! Answers was considerably worse than both. Similar experiments with two other set of questions confirmed the results.

References

[1] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

- [2] J. Graupmann, J. Cai, and R. Schenkel. Automatic query refinement using mined semantic relations. In *International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)*, Tokyo, Japan, 2005, pp. 205–213. IEEE. Acceptance ratio 1:4.

- [3] J. Graupmann and R. Schenkel. The light-weight semantic web: Integrating information extraction and information retrieval for heterogeneous environments. In *SIGIR 2005 Workshop on Heterogeneous and Distributed Information Retrieval (HDIR)*, Salvador, Brazil, August 2005, pp. 1–8. ACM.

[2] The witnesses could also be weighted by their authority, e.g. Page Rank.

- [4] J. Graupmann and R. Schenkel. GeoSphereSearch: Context-aware geographic web search. In R. Purves and C. Jones, eds., *3rd Workshop on Geographic Information Retrieval*, Seattle, WA, USA, 2006, pp. 1–4.
- [5] 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.

[6] 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.

- [7] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. Research Report MPI-I-2007-5-001, Max-Planck-Institut for Informatics, Saarbrücken, Germany, March 2007.
- [8] R. Schenkel, A. Theobald, and G. Weikum. Semantic similarity search on semistructured data with the XXL search engine. *Information Retrieval*, 8(4):521–545, December 2005.

16.4.5 Ontology Support

Investigator: Fabian Suchanek in collaboration with Peter Baumgartner

Recent years have seen an increasing interest in formal knowledge bases (KBs). Yet demanding application areas – notably the *Semantic Web* – will have to remain a vision without powerful automated reasoning support for these KBs. All reasoning tasks on KBs can be reduced to the problem of determining whether the KB is satisfiable or not. Unfortunately, this problem is undecidable for first-order logic KBs. Current off-the-shelf theorem provers generally perform well on determining the unsatisfiability of a KB. However, they often do not terminate when given a *satisfiable* KB[3]. To address this problem, we propose to use *model computation systems* instead of the standard theorem provers. These systems work bottom-up by computing all literals that can be derived from a set of logic formulae. Examples of such systems include dlv [2], smodels [3] and KRHyper [5]. As input, these systems require a *Disjunctive Logic Program (DLP)*. There exists already a well-known transformation from first order logic formulae to DLPs, which preserves satisfiability. We improve this naive transformation in several ways to make it suitable for ontological KBs[1].

First, our approach transforms away equality. This is crucial, because model generation systems usually do not have built-in equality handling. Second, our transformation can be configured to respect a certain form of the Unique Name Assumption (UNA). This is useful in the context of ontologies, when different constants are best considered to denote different objects. Third, our approach allows to avoid unnecessary Skolem terms: If an existentially quantified role is already filled, the existential formula is satisfied by that role-filler instead of creating a new Skolem term. This keeps the resulting models slim and meaningful. Third, our transformation allows a non-standard reading of existentially quantified formulas, namely as *integrity constraints*. That is, the model building process can be instructed to fail if an existentially quantified formula is not already fulfilled by the KB. Finally, our transformation allows a "loop check". This check can avoid infinite models in some cases by detecting finite

[3] see http://www.w3.org/2003/08/owl-systems/test-results-out for an exhaustive comparison of provers.

ones, i.e. the prover will terminate on certain satisfiable KBs even if it would not do so with the naive transformation.

We conducted preliminary experiments on the first-order portion of the SUMO ontology[4], one of the largest formal ontologies publicly available today. We showed that our model generation system KRHyper terminates on our transformation and that unnecessary Skolem terms are avoided.

References

- [1] P. Baumgartner and F. M. Suchanek. Automated reasoning support for first order ontologies. In J. J. Alferes, J. Bailey, W. May, and U. Schwertel, eds., *Principles and Practice of Semantic Web Reasoning, 4th International Workshop, PPSWR 2006*, Budva, Montenegro, 2006, *LNCS 4187*, pp. 18–32. Springer. Acceptance ratio 1:5.

[2] T. Eiter, W. Faber, N. Leone, and G. Pfeifer. Declarative problem-solving using the DLV system. In *Logic-based artificial intelligence*, pp. 79–103. Kluwer Academic Publishers, 2000.

[3] I. Niemelä and P. Simons. Efficient implementation of the well-founded and stable model semantics. In *Joint International Conference and Syposium on Logic Programming*, 1996, pp. 289–303. The MITPress.

[4] I. Niles and A. Pease. Towards a standard upper ontology. In *2nd International Conference on Formal Ontology in Information Systems, (FOIS)*, 2001, pp. 2–9.

[5] C. Wernhard. System Description: KRHyper. In *CADE-19 Workshop on Model Computation*, 2003.

16.4.6 Ranking for structured queries

Investigators: Gerhard Weikum in collaboration with Surajit Chaudhuri, Gautam Das, and Vagelis Hristidis

Even in structured databases that are searched with a precise query language like SQL, there is a need for ranking query results when too many answers are returned that all satisfy the specified query predicates. This situation typically arise in searching product catalogs for real estate, travel offers, etc., but it can be generalized too many other settings. For example, consider a potential home buyer who runs a query with a not very selective condition such as *City = Seattle and View = Waterfront*; this may result in too many tuples in the answer, since there are many homes with waterfront views in Seattle.

We have developed a novel model for ranking the results of such structured queries, based on data as well as workload statistics. Our approach builds on the Probabilistic Information retrieval paradigm. In estimating the probability of a record that satisfies the specified query conditions being relevant, we consider correlations between specified and unspecified attributes. For example, a query with predicate *SchoolQuality = high* can automatically prefer results with *View = Waterfront* even if the View attribute is not specified in the query, since these two attributes are highly correlated in the data. Furthermore, we consider functional dependencies among attributes, and we exploit correlations that are expressed in the workload rather than the data. For example, when users often query for homes with an *IN (Bellevue, Kirkland, Redmond)* predicate, we can infer that these three cities are similar to each other. Otherwise, we make the standard assumption of limited independence among

attributes for tractability. All this information is fed into the probabilistic relevance model, and produces much better rankings than a straightforward tf*idf approach would deliver. For efficient query evaluation, we precompute appropriate statistics and employ the threshold algorithm at run-time.

References

[1] S. Chaudhuri, G. Das, V. Hristidis, and G. Weikum. Probabilistic ranking of database query results. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, eds., *Proceedings 2004 VLDB Conference, The 30th International Conference on Very Large Databases (VLDB)*, Toronto, Canada, 2004, pp. 888–899. Morgan Kaufmann. Acceptance ratio 1:6.

- [2] S. Chaudhuri, G. Das, V. Hristidis, and G. Weikum. Probabilistic information retrieval approach for ranking of database query results. *ACM Transactions on Database Systems*, 31(3):1134–1168, September 2006.

- [3] S. Chaudhuri, R. Ramakrishnan, and G. Weikum. Integrating DB and IR technologies: What is the sound of one hand clapping? In M. Stonebraker, G. Weikum, and D. DeWitt, eds., *Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR 05)*, Asilomar, CA, USA, 2005, pp. 1–12. VLDB. Acceptance ratio 1:3.

16.5 Peer-to-Peer Search and Information Management

Coordinator: Christos Tryfonopoulos

16.5.1 P2P Web Search

Investigators: Matthias Bender, Tom Crecelius, Mouna Kacimi, and Sebastian Michel

Data on the Web is originally highly distributed, residing on millions of sites with more and more individuals contributing to it by, e.g., their blogs or personal web pages. This huge amount of information is currently managed by centralized architectures that reside on the computational power of large server farms. With the proliferation of P2P computing [11] to applications such as file sharing and IP telephony, various projects [5, 6, 13, 14], including our own Minerva project [1, 2, 3, 4, 7, 8], have started to build and operate P2P Web search systems that aim at harnessing the computational power residing "at the edges of internet". Additionally, the growing concern about centralized search engines and their susceptibility to commercial interests, spam or distortion by spam combat, biases in geographic and thematic coverage, or even censorship, has fueled this move towards solutions based on the P2P paradigm that can facilitate pluralism in a distributed and self-organizing way.

P2P computing is an intriguing paradigm for Web search and at the same time, it exposes large overlap with many traditional research areas (e.g., database systems, information retrieval, etc.) and can highly benefit from existing work. However, the special characteristics of P2P architectures (e.g., scale, dynamics, selfishness) require a different perspective on important design facets. The crucial challenge in developing a successful P2P Web search engine is based on reconciling the following conflicting goals: retrieving *high quality results* while ensuring *scalability* and *efficiency* in the presence of very large peer populations and high dynamics.

In the light of the above, we put forward our fully implemented P2P Web search prototype Minerva [1, 2]. Minerva consists of fully autonomous peers, each holding a local document collection (acquired by autonomous Web crawling, by importing content from external sources, or even by wrapping the contents of digital libraries [4]). It relies on a conceptually global but physically distributed directory which is layered on top of a Distributed Hash Table (DHT). DHTs [9, 10, 12] are second generation structured overlay networks devised as a remedy for efficiently solving the object location problem. Each Minerva peer is responsible for managing compact, aggregated information about the other peers' local knowledge. The DHT is used to partition the term space, such that every peer is responsible for the statistics and metadata of a randomized subset of terms within the directory [8]. To achieve this, each peer posts its own per-term summaries (called *Posts*) of its local data collection to the directory. Then, the DHT determines the peer currently responsible for this term. The corresponding peer maintains a *PeerList* of all postings for this term from across the network. Posts contain contact information about the peer who posted this summary together with statistics to calculate IR-style measures for a term. These statistics are used to support the query process, i.e., determining the most promising peers for a particular query [3, 7]. To reduce the load on the distributed directory and, thus, to support an a-priori unlimited number of peers, only statistical features that characterize a peer's specific content and focus are posted to the directory. As most peers are distinguished authorities only for a few topics, this posting strategy drastically reduces the storage and bandwidth load necessary for directory maintenance.

Each peer may autonomously run a search engine on a personalized local corpus (e.g., built from a thematically focused Web crawl) and collaborate with other peers to receive further or improved results. In this way, a P2P search engine would exploit techniques such as collaborative search and recommendations, and can potentially benefit from the intellectual input (e.g., tagging, bookmarks, etc.) of a large user community that participates in the data sharing network [1]. Additionally the vast processing power of a peer-to-peer network, could enable much more advanced methods for linguistic data analysis, statistical learning, or ontology-based background knowledge and reasoning.

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 07)*, Asilomar, USA, January 2007, pp. 164–168. www.crdrdb.org.

- [2] 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.

- [3] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. P2P content search: Give the web back to the people. In *5th International Workshop on Peer-to-Peer Systems (IPTPS 2006)*, Santa Barbara, US, 2006. Acceptance ratio: 1:4.

- [4] M. Bender, S. Michel, C. Zimmer, and G. Weikum. Towards collaborative search in digital libraries using peer-to-peer technology. In C. Türker, M. Agosti, and H.-J. Schek, eds., *Peer-to-peer, grid, and service-orientation in digital library architectures, 6th Thematic Workshop of*

the EU Network of Excellence DELOS, Cagliari, Italy, August 2005, *LNCS 3664*, pp. 80–95. Springer. Selected, Revised Papers.

[5] F. M. Cuenca-Acuna, C. Peery, R. P. Martin, and T. D. Nguyen. PlanetP: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities. In *Proceedings of the International Symposium on High Performance Distributed Computing (HPDC)*, 2003.

[6] J. Lu and J. Callan. Content-Based Retrieval in Hybrid Peer-to-Peer Networks. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, 2003, pp. 199–206. ACM Press.

- [7] S. Michel, M. Bender, P. Triantafillou, G. Weikum, and C. Zimmer. P2P web search with MINERVA: How do you want to search tomorrow? (demo), 2005.
- [8] S. Michel, P. Triantafillou, and G. Weikum. Minerva∞: A scalable efficient peer-to-peer search engine. In G. Alonso, ed., *Middleware 2005, ACM, IFIP, USENIX 6th International Middleware Conference*, Grenoble, France, 2005, *LNCS 3790*, pp. 60–81. Springer.

[9] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A Scalable Content-addressable Network. In *Proccedings of teh ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM)*, 2001.

[10] A. Rowstron and P. Druschel. Pastry: Scalable, Decentralised Object Location and Routing for Large-Scale Peer-to-Peer Systems. In *Proceedings of the International Conference on Distributed Systems Paltforms (Middleware)*, November 2001.

[11] R. Steinmetz and K. Wehrle, eds. *Peer-to-Peer Systems and Applications*, *LNCS 3485*. Springer, 2005.

[12] I. Stoica, R. Morris, D. Liben-Nowell, D. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan. Chord: a Scalable Peer-to-Peer Lookup Protocol for Internet Applications. *IEEE/ACM Transactions on Networking*, 11(1):17–32, 2003.

[13] C. Tang, Z. Xu, and M. Mahalingam. pSearch: Information Retrieval in Structured Overlays. In *Proceedings of the International Workshop on Hot Topics in Networks (HotNets)*, 2002.

[14] C. Tryfonopoulos, S. Idreos, and M. Koubarakis. LibraRing: An Architecture for Distributed Digital Libraries Based on DHTs. In *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Vienna, Austria, September 2005, pp. 25–36.

16.5.2 Query Routing

Investigators: Matthias Bender, Sebastian Michel, Josiane Parreira, Gerhard Weikum, and Christian Zimmer in collaboration with Peter Triantafillou and Nikos Ntarmos

One of the fundamental functionalities that a P2P Web search system must provide is to identify the most "appropriate" peers for a particular query, i.e., those peers that are expected to locally hold high-quality results for the query should be involved in the query processing. This task is commonly referred to as *query routing*, sometimes also as *resource* or *collection selection*.

We stress that query routing is more challenging than it may appear at first sight: the set of peers to be contacted is not simply the set of all peers that store relevant index data. Such a set could contain a very large number of peers and contacting all of them is prohibitive.

While there exist a number of approaches for query routing in the literature on distributed IR – e.g., CORI [9], GlOSS [12], and methods based on statistical language models [18]

– these were typically designed for a stable and rather small set of collections (e.g., in the context of metasearch engines). Our experiments have shown that they fall short of addressing the peculiarities of a large-scale, highly dynamic P2P system [5]. Also, they disregard personalization as an integral potential of a distributed system, which we have addressed in the context of digital libraries [2, 1]. A second problem which we have addressed in the context of query routing is the process of merging the results obtained from remote sources [10].

Overlap-aware routing.

A key shortcoming of the existing strategies is their – at first sight by all means intuitive – approach to base their decisions on the expected result quality of the peers. However, with autonomous peers harvesting information at their own discretion, peers may have highly overlapping local data contents as popular information is indexed by a large number of peers individually.

Thus, the rationale for the overlap-aware query routing strategies is based on the observation that a query should be forwarded to peers that are expected to contribute not only high-quality, but also *complementary results*. If a remote peer returns more or less the same high-quality results that the query initiator already obtained from other candidates, then the whole approach of collaborative P2P search would be pointless. An integrated quality- and overlap-aware query routing method needs to be able to estimate the content richness of candidate target peers, in terms of the similarity and relevance of the peers' contents to the given query, and the degree of novelty that a candidate peer would offer relative to the initial results that are already known to the query originator.

From a research perspective, the key challenges in order to achieve overlap-awareness lie in:

- defining appropriate *metrics* that allow an estimation of the expected benefit that the inclusion of a peer will bring to the result set,
- representing local index data by means of compact synopses that support operations to estimate the above metrics for a given collection, but also support *aggregation* of synopses for several collections, and
- designing scalable *algorithms* for overlap-aware query routing in P2P information systems that can benefit from these synopses efficiently.

We have addressed the first issue by introducing the notion of *novelty* that combines standard set operations (union, intersection, cardinality) as a metric for the contribution of a collection with regard to a given reference collection. We have evaluated a number of statistical synopses (Bloom filters [6], hash-sketches [11], min-wise independent permutations [8]) from the literature regarding their general accuracy and their particular support for the necessary intermediate aggregation steps. We have presented the IQN algorithm (Integrated Quality and Novelty) [14] that chooses target peers in an iterative manner, performing two steps in each iteration: first, the Select-Best-Peer step identifies the most promising peer regarding a combination of result quality and novelty. This step is driven by the statistical

synopses that are obtained from the distributed metadata directory. Then, the Aggregate-Synopses step conceptually aggregates the selected peer's content with the previously selected peers' data collections. This aggregation is actually carried out on the compact synopses, not on the full data. The two-step selection procedure is iterated until given performance and/or quality goals are satisfied (e.g., a predefined number of peers is reached, or a desired recall is estimated to be achieved).

Our experiments have shown that the efficiency and effectiveness of the IQN routing method crucially depends on appropriately designed compact synopses describing the collection of a peer. The synopses must be small to keep network bandwidth consumption and storage costs low, yet they must offer low-error estimations of quality and novelty measures. Furthermore, to support the intermediary aggregation step introduced above, it must be possible to iteratively combine multiple synopses published by different peers in order to derive a synopsis for a hypothetical combined collection. We have developed such methods using Bloom filters, hash sketches, and min-wise independent permutations as peer synopses [3, 14]. Extensive experiments have shown that these novel algorithms indeed combine very low overhead with high accuracy for quality-novelty estimation, and the IQN query routing strategy outperforms prior methods for query routing in our setting.

We have also developed the novel notion of *global document occurrences* (GDO) [15] that, when processing a query, penalizes frequent documents increasingly as more and more peers contribute their local results by adjusting their local contribution to the relevance scores for query routing. This is orthogonal to the overlap-aware query routing proposed earlier: while the previous methods use statistical synopses (e.g., Bloom filters) describing the local indexes to estimate the cardinality of mutual overlap between the peers, the GDO approach takes into account the frequency of individual documents and adjusts their relevance scores accordingly.

Correlation-aware routing.

The summaries describing a peer in the distributed metadata directory are usually organized on a per-term basis, indicating the expected result quality of a peer's collection for a given term. This limitation is considered unavoidable, as statistics on all term pairs would incur a quadratic explosion, leading to a breach with the goal of scalability. On the other hand, completely disregarding correlations among terms is a major impediment: for example, consider the following extreme scenario. Assume peer p_1 contains a large number of data items for each of the two terms a and b separately, but none that contains both a and b together. Judging only by per-term statistics, state-of-the-art query routing approaches would reach the conclusion that p_1 is a good candidate peer for the query $(\{a, b\}, k)$, whereas the actual result set would be empty. This is because the summaries describe the expected quality for each term *separately*, where the most promising candidate peers for the entire query should really exhibit a higher-than-average frequency of data items that contain both terms *at the same time*.

We have developed and evaluated two approaches to address this issue [13]:

- *sk-STAT* estimates the desired multi-term statistics from the existing per-term statistics with additional computational efforts and at higher networking costs, and

– *mk-STAT* enhances the distributed directory to explicitly include also statistical information about judiciously chosen sets of multiple terms

The caveat of mk-STAT is that it faces the necessity to identify those valuable term sets to avoid the dimension explosion mentioned above. It does so by mining locally gathered query logs, to improve the performance of frequently queried term combinations. This discovery phase additionally performs an in-depth statistical analysis of the degree of correlation observed within the peers' data collections for these term combinations. One of our novel contributions has been how to make this analysis efficient and scalable.

sk-STAT, on the other hand, can readily deal with all possible term sets, as it only relies on combinatorial operations on the existing single-term statistics. However, it has higher bandwidth requirements at query time, as larger amounts of these single-keyword statistics have to be shipped to estimate the statistics for term sets.

Authority-aware routing.

Authority measures, such as the popular PageRank algorithm [7], are a powerful technique to improve the result ranking quality in centralized Web search. Its intuitive approach to favor documents from high-authority sources has been proven a pivotal ingredient to sophisticated document scoring models, beyond popular statistical information describing the documents.

However, the computation of authority scores has traditionally required complete knowledge of the underlying Web graph, i.e., knowledge of all documents (nodes) and interconnecting hyperlinks (edges), whereas we envision a distributed system of autonomous peers that independently crawl the Web, effectively leading to (typically overlapping) partitions of the Web graph being stored at the peers. We proposed JXP, the first solution to efficiently compute authority scores that provably converge to global PageRank authority scores in such an environment in a completely decentralized manner [16] (see Section 16.5.5).

Conceptually, the query routing process closely resembles the local document scoring process when evaluating a query. If we visualize each peer as one large *superdocument* (the combination of all its local documents) like query routing strategies based on statistical language models, query routing boils down to finding the top-k "documents" in the network. As PageRank authority scores have been shown to greatly improve this process, it suggests itself to exploit global PageRank scores for query routing purposes, so to base query routing on an a scoring model that better captures the expected result quality of a peer.

The key idea is to improve the query routing process by preferring peers that have a high local PageRank score mass. We have developed different approaches that take into account the total PageRank score mass of a collection, query-specific portions of the PageRank score mass, and a hybrid framework that combines the authority score mass with other existing scoring models for query routing [17], such as CORI. We have discussed the impact that both ingredients have on query routing and have provided evidence gained from experiments that shows the impact on query result quality.

References

• [1] M. Bender, N. Fuhr, Y. Ioannidis, D. Kossman, H.-J. Schek, G. Weikum, P. Zezula, and C. Zimmer. Personalized query routing in peer-to-peer federations of digital libraries. In C. Thanos,

ed., *DELOS Research Activities 2006*, pp. 29–32. IST-CNT, Pisa, Italy, 2007.

- [2] M. Bender, Y. Ioannidis, H. Nottelmann, H.-J. Schek, G. Weikum, P. Zezula, and C. Zimmer. Personalized query routing in peer-to-peer federations of digital libraries. In C. Thanos, ed., *DELOS Research Activities 2005*, pp. 17–18. ISTI-CNR, Pisa, Italy, 2007.

- [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 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brazil, 2005, pp. 67–74. ACM. Acceptance ratio 1:5.

- [4] M. Bender, S. Michel, G. Weikum, and C. Zimmer. Das MINERVA-Projekt: Datenbankselektion für Peer-to-Peer-Websuche. *Informatik - Forschung und Entwicklung*, 20(3):152 – 166, December 2005.

- [5] M. Bender, S. Michel, G. Weikum, and C. Zimmer. The MINERVA project: Database selection in the context of P2P search. In G. Vossen, F. Leymann, P. C. Lockemann, and W. Stucky, eds., *Datenbanksysteme in Business, Technologie und Web (BTW), 11. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS)*, Karlsruhe, Germany, March 2005, *Lecture Notes in Informatics*, vol. 65, pp. 125–144. Gesellschaft für Informatik. Acceptance ratio 1:3.

[6] B. H. Bloom. Space/Time Trade-offs in Hash Coding with Allowable Errors. *Communications of the ACM*, 13(7):422–426, 1970.

[7] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30(1-7):107–117, 1998.

[8] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher. Min-Wise Independent Permutations. *Journal of Computer and System Sciences (JCSS)*, 60(3):630–659, 2000.

[9] J. P. Callan, Z. Lu, and W. B. Croft. Searching Distributed Collections with Inference Networks. In E. A. Fox, P. Ingwersen, and R. Fidel, eds., *Proccedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, 1995, pp. 21–28. ACM Press.

- [10] S. Chernov, P. Serdyukov, M. Bender, S. Michel, G. Weikum, and C. Zimmer. Database selection and result merging in P2P web search. In *3rd International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2005)*, Trondheim, Norway, 2005, pp. 1–14. Acceptance rate 1:3.

[11] P. Flajolet and G. N. Martin. Probabilistic Counting Algorithms for Data Base Applications. *Journal of Computer and System Sciences (JCSS)*, 31(2):182–209, 1985.

[12] L. Gravano, H. Garcia-Molina, and A. Tomasic. GlOSS: Text-Source Discovery over the Internet. *ACM Transactions of Database Systems*, 24(2):229–264, 1999.

- [13] 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.

- [14] S. Michel, M. Bender, P. Triantafillou, and G. Weikum. IQN routing: Integrating quality and novelty in P2P querying and ranking. In Y. Ioannidis, M. H. Scholl, J. W. Schmidt, F. Matthes, M. Hatzopoulos, K. Boehm, A. Kemper, T. Grust, and C. Boehm, eds., *Advances in Database Technology - EDBT 2006: 10th International Conference on Extending Database Technology*, Munich, Germany, March 2006, *LNCS 3896*, pp. 149–166. Springer.

- [15] O. Papapetrou, S. Michel, M. Bender, and G. Weikum. On the usage of global document occurrences in peer-to-peer information systems. In R. Meersman, Z. Tari, M.-S. Hacid, J. Mylopoulos, B. Pernici, Ö. Babaoglu, H.-A. Jacobsen, J. P. Loyall, M. Kifer, and S. Spaccapietra, eds., *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005*, Agia Napa, Cyprus, 2005, *LNCS 3760*, pp. 310–328. Springer.

- [16] 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.

- [17] J. X. Parreira, S. Michel, and M. Bender. Size doesn't always matter: Exploiting pagerank for query routing in distributed IR. In *P2PIR '06: Proceedings of the International Workshop on Information Retrieval in Peer-to-peer Networks*, Arlington, USA, 2006, pp. 25–32. ACM.

[18] L. Si, R. Jin, J. P. Callan, and P. Ogilvie. A Language Modeling Framework for Resource Selection and Results Merging. In *Proccedings of the International Conference on Information and Knowledge Management (CIKM)*, 2002, pp. 391–397. ACM.

16.5.3 P2P Statistics

Investigators: Matthias Bender, Sebastian Michel in collaboration with Peter Triantafillou

Given the large-scale data distribution in a P2P system like Minerva, one of the key technical challenges is *result merging*, i.e., the process of effectively combining local query results from different sources. While document scoring and ranking is a challenging problem already in centralized systems, additional difficulty in a distributed environment stems from the fact that most of the popular document scoring models, such as tf*idf or more advanced models like Okapi BM25, use collection-specific statistical information for this purpose. Most prominently, both use *document frequencies (df)*, i.e. the number of documents in the collection that contain a query term. Note the difference to the notion of *peer* or *collection frequencies* that estimate the number of *collections* that contain a query term. The *document frequency*, instead, represents the total number of distinct documents that contain a term. The local usage of collection-specific *df* values in these scoring models result in document scores that are incompatible across collections and, thus, make result merging difficult. On the other hand, if *global df* values could be applied, the document scoring and ranking would be ideal in the sense that it would be identical to the document ranking that would be produced by a hypothetical combined collection.

Early research on distributed information retrieval systems typically assumed disjointly partitioned collections. In such a setting, the *global df* value is simply the sum over all local *df* values. Instead, we envision autonomous peers that independently gather thematically focused collections through web crawls or similar techniques. In such a setting, studies show a skewed distribution of documents across the collections, with popular documents contained in a large fraction of collections. Thus, summing up the *df* values across collections would inevitably lead to biased *df* values (and, thus, document scores), as popular documents are repeatedly accounted for. Additionally, thematically focused collections show a high variance of *df* values for the same term (whereas randomly partitioned collections show a rather

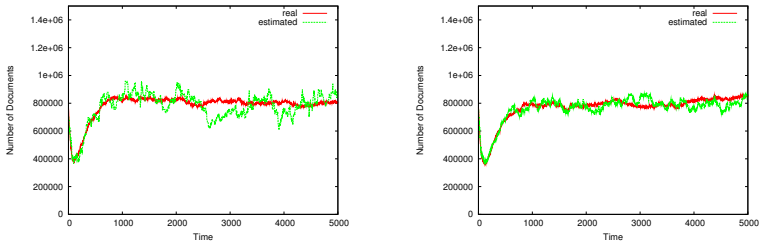


Figure 16.2: *df* Estimation Accuracy using 64-bitmap (left) or 256-bitmap (right) Hash-Sketches

uniform distribution of *df* values for the same term). This further increases the necessity of a score normalization across peers.

In [1], we present a robust and scalable approach towards estimating *global df* values using hash sketches [2]. We study the general accuracy of hash sketches when used as synopses to estimate document frequencies and we develop an efficient strategy to combine these hash sketch synopses across collections in a way that does not incur any additional error from combining them. The proposed technique can be directly applied in systems like Minerva that maintain a term-based directory. Every peer creates a hash sketch synopsis for each term, i.e., each hash sketch describes the documents that contain a particular term. When publishing meta data to the directory, peers include these synopses. The directory peer for term a is then able to determine the global number of distinct documents that contain term a by inspecting the sketches.

Figure 16.2 shows the result of our experiment where we consider the accuracy of the global document frequency estimation under churn, i.e., peers entering and leaving the network. The x-axis shows the evolving time and the y-axis the actual number of documents given by peers currently in the system versus the estimated number of documents in the current system. Although there is a lot of churn, in particular in the beginning of the experiment, the hash sketch based technique performs remarkably well.

Furthermore, we show the superiority of our *global df* estimation technique compared to other techniques and present experimental evidence of the effectiveness improvements in result merging stemming from this improved knowledge. The experiments are conducted on real-word web data using our fully operational P2P Web search engine prototype.

Although this approach works fine for our Minerva system, there are some concerns with respect to load imbalances, i.e., a peer responsible for collecting the hash sketch for a very popular term will receive a huge number of requests. To overcome these limitations, we have developed *Distributed Hash Sketches*, a truly decentralized counting technique [4]. The main idea is to distribute a single hash sketch for a particular attribute (e.g. term) across multiple peers. Hence, the load of maintaining this hash sketch is distributed.

References

- [1] M. Bender, S. Michel, P. Triantafillou, and G. Weikum. Global document frequency estimation in peer-to-peer web search. In D. Zhou, ed., *9th International Workshop on the Web and Databases (WebDB 2006) @ SIGMOD2006*, Chicago, USA, 2006, pp. 69–74. Acceptance Ratio 1:4.

[2] P. Flajolet and G. N. Martin. Probabilistic Counting Algorithms for Data Base Applications. *Journal of Computer and System Sciences (JCSS)*, 31(2):182–209, 1985.

- [3] A. Linari and G. Weikum. Efficient peer-to-peer semantic overlay networks based on statistical language models. In *P2PIR '06: Proceedings of the International Workshop on Information Retrieval in Peer-to-peer Networks*, Arlington, Virginia, USA, 2006, pp. 9–16. ACM.

- [4] N. Ntarmos, P. Triantafillou, and G. Weikum. Counting at large: Efficient cardinality estimation in internet-scale data networks. In L. Liu, A. Reuter, K.-Y. Whang, and J. Zhang, eds., *Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006)*, Atlanta, GA, USA, 2006, pp. 1–10. IEEE.

16.5.4 P2P Information Filtering

Investigators: Christos Tryfonopoulos and Christian Zimmer in collaboration with Manolis Koubarakis

In a highly dynamic environment such as the Web, where users have to sift through enormous amounts of new information to satisfy an information demand, *information filtering* (IF), also referred to as *publish/subscribe* or *continuous querying*, is of equal importance to one-time querying. In IF users are able to subscribe to information sources and be notified when new documents of interest are published. In such a scenario, a user submits a subscription (also known as continuous query or profile) to the system to receive notifications whenever certain events of interest take place (e.g., a document on the topic of interest becomes available).

In this context, we put forward MAPS (Minerva Approximate Publish/Subscribe) [2], an architecture to support approximate information filtering functionality in a distributed P2P environment. Most existing information filtering approaches [4, 5, 6] have the underlying hypothesis of potentially delivering notifications from every information source. The MAPS approach relaxes this assumption by monitoring only selected sources that are likely to publish documents relevant to a user interest in the near future. In MAPS, a user subscribes with a continuous query and monitors the most interesting information sources in the network. These selected sources store the subscription locally and only published documents from these sources are forwarded to the user. In this way, MAPS enhances scalability by trading recall for lower message traffic. A brief comparison between exact and approximate IF can be found in [1].

Obviously, the most critical decision in MAPS denotes the selection of the most promising information sources. To achieve this, MAPS uses a Minerva-style distributed directory that maintains statistical information about peers' publications by utilizing a DHT. Contrary to the search scenario, resource selection techniques (e.g., CORI [3]) are not sufficient such that the publishing behavior of information sources has to be considered. The source selection approach of MAPS combines resource selection with publishing behavior of IR statistics using time-series analysis with double exponential smoothing techniques. This way, MAPS recognizes new information sources currently publishing many relevant documents to a subscription instead of huge collections with low publishing rates.

The experimental evaluation considered different publishing scenarios and showed that approximate information filtering is a promising research direction to improve the scalability of P2P IF. In some cases, an average recall as high as 80% can be achieved by subscrib-

ing a continuous query to only 8% of the information sources such that network traffic at publication time is extremely reduced by spending more effort at query indexing time.

References

- [1] M. Bender, S. Michel, S. Parkitny, and G. Weikum. A comparative study of pub/sub methods in structured P2P networks (to be published). In S. Bergamaschi, S. Joseph, J.-H. Morin, and G. Moro, eds., *Fourth International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2006)*, Seoul, South Korea, 2006. Also published at Delos.
- [2] K. Berberich, M. Koubarakis, C. Tryfonopoulos, G. Weikum, and C. Zimmer. MAPS: Approximate publish/subscribe functionality in peer-to-peer networks. In *ADPUC '06: Proceedings of the 1st International Workshop on Advanced Data Processing in Ubiquitous Computing (ADPUC 2006)*, Melbourne, Australia, 2006, *ACM International Conference Proceeding Series*, vol. 181, pp. 1–6. ACM.

[3] L. Si, R. Jin, J. P. Callan, and P. Ogilvie. A Language Modeling Framework for Resource Selection and Results Merging. In *Proccedings of the International Conference on Information and Knowledge Management (CIKM)*, 2002, pp. 391–397. ACM.

[4] C. Tang and Z. Xu. pFilter: Global Information Filtering and Dissemination Using Structured Overlays. In *Proceedings of the International Workshop on Future Trends in Distributed Computing Systems (FTDCS)*, 2003.

[5] C. Tryfonopoulos, S. Idreos, and M. Koubarakis. LibraRing: An Architecture for Distributed Digital Libraries Based on DHTs. In *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Vienna, Austria, September 2005, pp. 25–36.

[6] C. Tryfonopoulos, S. Idreos, and M. Koubarakis. Publish/Subscribe Functionality in IR Environments Using Structured Overlay Networks. In *Proccedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, 2005.

16.5.5 P2P Link Analysis for Authority and Trust

Investigators: Josiane Parreira, Sebastian Michel, and Gerhard Weikum in collaboration with Debora Donato and Carlos Castillo

PageRank-style authority scoring, based on the Eigen-space analysis of a suitably defined graph of Web links, endorsements, or interactions, is an established tool for ranking information units (Web pages, sites, peers, social groups, etc.) by their relative importance. As Google has impressively demonstrated, such authority information can be exploited for improved ranking of search results.

Recently, various techniques have been proposed for distributed PageRank-style authority computations. However, these advanced methods work only when the overall Web graph is partitioned into disjoint fragments. With autonomous peers creating (or gathering) data at their own discretion, this is not a reasonable assumption.

JXP (Juxtaposed Approximate PageRank) [2, 3] is an algorithm for dynamically computing, in a decentralized P2P manner, global authority scores when the Web graph is spread across many autonomous peers with arbitrarily overlapping graph fragments and the peers are a priori unaware of other peers' fragments. Each peer periodically, and independently of other peers, performs local PageRank score computations on its local graph fragment, where

the local graph is augmented by a *world node* that represents the locally unknown part of the global graph. Mathematically, this is a state lumping or aggregation technique for the underlying Markov chain.

JXP initiates random meetings between pairs of peers, for mutual exchange of information about their local graph fragments and to continuously improve each peer's knowledge about its world node. The meetings take place asynchronously, without any central planning. The world node is constructed by combining all edges from local pages (or other kinds of graph nodes) that point to external (i.e., non-local) pages into edges to the world node. Analogously, when a peer learns (by means of meeting another peer) about an edge from a non-local page to a local page, a corresponding edge from the world node to that local page is added to the local graph. Additionally, the world node has a self-loop edge that represents all links among external documents.

Throughout these meetings, the JXP scores that are locally maintained at each peer for its graph fragment converge to the global authority scores that would be derived from the entire global graph. Figure 16.3 illustrates a meeting between two peers.

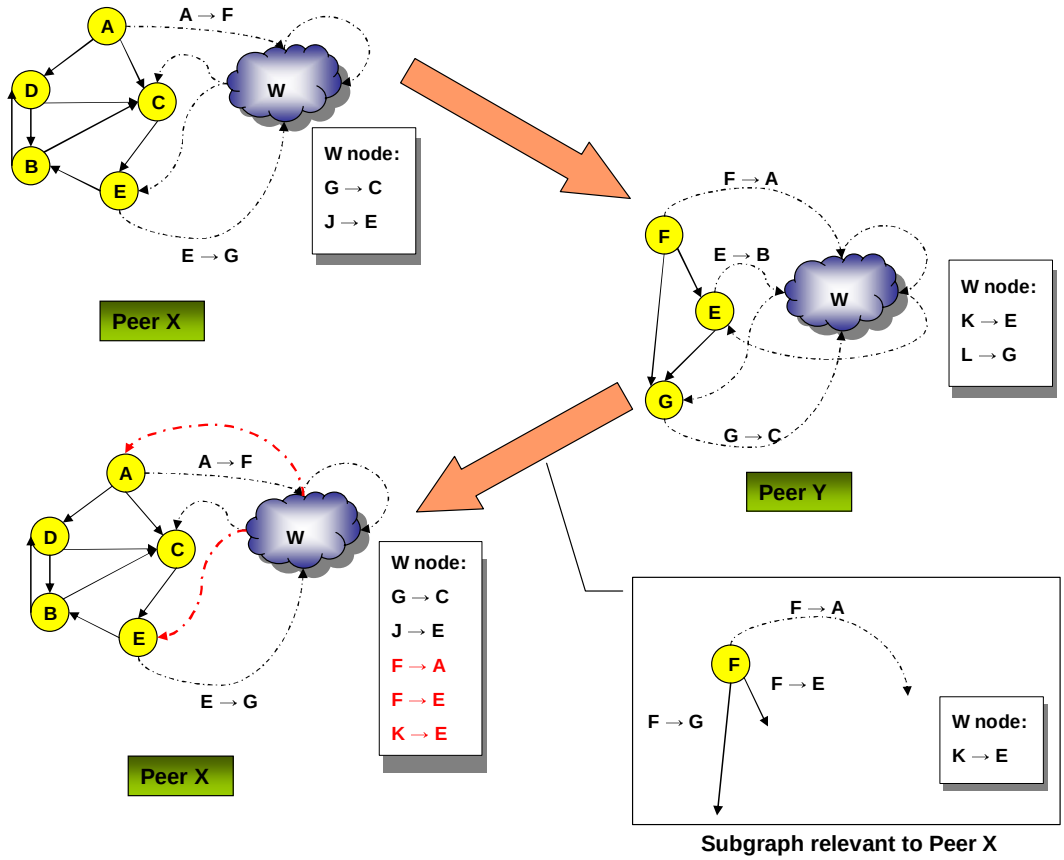


Figure 16.3: Improving Knowledge by Peer Meetings

The JXP algorithm is efficient and scalable, as all computations are strictly local and performed only on the small local graph fragments, whose sizes are independent of the

number of peers in the network or the size of the global graph. Network performance can be improved by a heuristic strategy for guiding the choice of peers for a meeting.

Since high authority scores can bring benefits for peers, it is expected that malicious peers would try to distort the correctness of the algorithm, by providing different (usually higher) scores for some of their local pages. P2P networks are generally vulnerable to malicious agents that can cheat in order to get more benefits. Therefore, P2P architectures for information sharing, search, and ranking must integrate a complete *reputation systems*. Reputation systems operate by collecting information on the behavior of the peers, scoring each peer based on good vs. bad behavior, and allowing the system to take countermeasures against suspicious peers.

We have also developed a trust model that integrates decentralized authority scoring with an equally decentralized reputation system. Our approach is based on anomaly detection techniques that allow us to detect a suspicious peer based on the deviation of its behavior from some common features that constitute the usual peer profile. Our method combines an analysis of the authority score distribution and a comparison of score rankings for a small set of pages. The JXP algorithm is then enhanced to avoid the impact of malicious peers. We call this enhanced version *TrustJXP* [1].

References

- [1] J. X. Parreira, C. Castillo, D. Donato, and G. Weikum. Computing trusted authority scores in peer-to-peer networks. Research Report DELIS-TR-460, University of Paderborn, Heinz Nixdorf Institute, Paderborn, Germany, 2006.
- [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 and G. Weikum. JXP: Global authority scores in a P2P network. In A. Doan, F. Neven, R. McCann, and G. Jan Bex, eds., *Proceedings of the 8th International Workshop on the Web & Databases (WebDB 2005) collocated with ACM SIGMOD/PODS 2005*, Baltimore, Maryland, USA, 2005, pp. 31–36. ACM.

16.5.6 Semantic Overlay Networks

Investigators: Josiane Parreira and Sebastian Michel in collaboration with Alessandro Linari

Semantic overlay networks (SONs) have lately been devised as a means to improve search effectiveness and efficiency in a P2P setting, while maintaining high peer autonomy and avoiding expensive routing table maintenance protocols necessary for structured overlays. In recent proposals of SONs [1, 5, 2], peers are connected to other peers of similar interests based on a rigid semantic profile. This restricts peer autonomy as it suggests clustering peers in specific groups and making networks that are not able to follow the dynamics that are introduced by changes in these interests. To address this we propose a strategy that follows the spirit of peer autonomy and creates semantic overlay networks based on the notion of

"peer-to-peer dating" [4]. Peers are free to decide which connections they create and which they want to avoid based on various usefulness estimators. The proposed techniques can be easily integrated into existing systems as they require only small additional bandwidth consumption as most messages can be piggybacked onto established communication. These additional semantic relations can be of great benefit during query routing in search engines, such as Minerva.

In [3] we consider a similar approach but instead of using the non-metric Kullback-Leibler divergence to compute the similarity between them, we use a symmetrized and "metricized" related measure, the square root of the Jensen-Shannon divergence, which let us map the problem to a metric search problem. The search strategy exploits the triangular inequality to efficiently prune the search space and relies on a priority queue to visit the most promising peers first. To keep communications costs low and to perform an efficient comparison between Language Models, we devise a compression technique that builds on Bloom-filters and histograms and we provide error bounds for the approximation and a cost analysis for the algorithms used to build and maintain the SON.

References

[1] K. Aberer, P. Cudré-Mauroux, M. Hauswirth, and T. V. Pelt. GridVine: Building Internet-Scale Semantic Overlay Networks. In *Proccedings of the International Semantic Web Conference (ISWC)*, 2004, pp. 107–121.

[2] M. Bawa, G. Manku, and P. Raghavan. SETS: Search Enhanced by Topic Segmentation. In *Proccedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*, 2003, pp. 306–313.

- [3] A. Linari and G. Weikum. Efficient peer-to-peer semantic overlay networks based on statistical language models. In *P2PIR '06: Proceedings of the International Workshop on Information Retrieval in Peer-to-peer Networks*, Arlington, Virginia, USA, 2006, pp. 9–16. ACM.

- [4] J. X. Parreira, S. Michel, and G. Weikum. p2pDating: Real life inspired semantic overlay networks for web search. *Information Processing & Management*, 43(3):643–664, 2007.

[5] P. Triantafillou, C. Xiruhaki, M. Koubarakis, and N. Ntarmos. Towards High Performance Peer-to-Peer Content and Resource Sharing Systems. In *Proceedings of the Biennial Conference on Innovative Data Systems Research (CIDR)*, 2003.

16.5.7 Benchmarks for P2P Retrieval

Investigators: Matthias Bender, Sebastian Michel, and Thomas Neumann

The design of a large-scale distributed search engine that aims deployment at thousands of nodes involves extensive experimentation and performance evaluation that will reveal possible weaknesses and allow for their early confrontation. Unfortunately, there is no standard benchmark that could be used for this task, resulting in every paper considering a different benchmark for its experimental evaluation. This fact renders comparison across different solutions and quantification of performance improvements an impossible task. In [1] we have presented a standardized, general purpose benchmark for Peer-to-Peer information retrieval system. The proposed benchmark is based on Wikipedia articles. However, the main problem

when designing a benchmark for distributed systems is not the choice of the underlying data collection but the assignment of documents to peers.

We use a graph-based clustering strategy that clusters the data such that the number of edges between clusters in minimized, which means that data linking to each other is placed in the same cluster. This is a reasonable clustering that is a good approximation to a crawler behavior. Unfortunately, finding the optimal clustering is NP-hard and the existing clustering algorithms usually violate some of our requirements. Therefore we propose a simple greedy cluster algorithm that can be used to partition the data suitably.

The documents in each topic (cluster) are then clustered again into smaller related chunks, using the same clustering algorithm on each topic cluster. These *chunks* are assigned to peers using a sliding window technique: Each peer is assigned a certain number of chunks and the next peer shifts the window a certain number of steps, creating any desired degree of overlap. As for the query workload, we propose the usage of a few most popular queries taken from Google's Zeitgeist query collection.

References

- [1] T. Neumann, M. Bender, S. Michel, and G. Weikum. A reproducible benchmark for P2P retrieval. In P. Bonnet and I. Manolescu, eds., *Proceedings of the 1st International Workshop on Performance and Evaluation of Data Management Systems, ExpDB 2006, in cooperation with ACM SIGMOD*, Chicago, Illinois, USA, 2006, pp. 1–8. ACM.

16.5.8 P2P Classification and Clustering

Investigators: Stefan Siersdorfer and Sergej Sizov

Another interesting problem in P2P networks that is well understood in centralized settings, e.g. in the context of Web retrieval, is text processing using machine learning algorithms (e.g. using supervised methods such as classification or unsupervised algorithms like clustering). In the context of a P2P network that puts together multiple users with shared topics of interest, it is natural to aggregate their knowledge and construct better machine learning models that could be used by every network member.

The naive solution would be to share available data (training samples and/or results of the focused crawl) along a higher number of peers with others. However, the following reasons may prevent the peer from sharing all of its data with other members of the overlay network:

- significantly increased network costs for downloads of additional training data on every peer
- increased runtimes for the training of the decision models
- privacy, security, and copyright aspects of the user's personal informations sources

To overcome the limitations of single-peer models, we propose the application of meta methods [1]. Our objective is to combine multiple independently learned models from several peers and to construct the advanced decision model that utilizes the knowledge of multiple P2P users. In addition we show how meta learning can be applied in a restrictive manner,

i.e., leaving out some documents rather than assigning them to inappropriate topics or clusters with low confidence, providing us with significantly more accurate classification and clustering results on the remaining documents.

References

- [1] S. Siersdorfer and S. Sizov. Automatic document organization in a P2P environment. 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. 265–276. Springer.

16.5.9 Failure Masking in Composite Web Services

Investigators: German Shegalov and Gerhard Weikum in collaboration with Roger Barga and David Lomet

A growing number of Internet businesses deliver mission-critical applications (stock trading, auctions, peer-to-peer information sharing and collaboration, etc.) to their customers as Web Services. These applications comprise heterogeneous components distributed over multiple layers. They pose strong requirements for service and consistent-data availability. Since many systems count many millions of lines of code, some bugs pass quality assurance undetected which leads to unpredictable service outages at some point. Recovery in transactional systems, as supported by the backend data servers, guarantees atomicity and persistence, but incorrect failure handling in applications often leads to incomplete or to unintentional *non-idempotent* request executions. This happens because many applications that are stateful by nature, i.e., with a state remembered between consecutive interactions, are rendered stateless, where all interactions are independent for easier manageability. Consequently, timeouts and resent messages among application servers or Web Services may lead to unintended duplication effects such as delivering two tickets for a single-ticket purchase request, resulting in severe customer irritation and business losses. The standard solution in the TP and DBMS world requires all state information of the application to be managed in the database or in transactional queues, but this entails a specific programming discipline that is often viewed as an unnatural burden by application developers.

To overcome these problems, we have developed the *Interaction Contracts (IC)* framework, providing a generic solution by means of integrated data, process, and message recovery. IC's *mask* failures, and allows programmers to concentrate on the application logic, greatly simplifying and speeding up application development. A challenge in implementing this kind of comprehensive multi-tier application recovery is to minimize the overhead of synchronous disk writes forced by the recoverability criteria.

More recently, we have developed the *Exactly-Once Web Service* platform *EOS*^{2} for composite Web Services with recovery guarantees. The EOS^2 implementation masks failures by adding a recovery layer to popular Web technology products: (i) the server-side script language PHP run on Apache Web server, and (ii) Internet browsers like IE to deliver recovery guarantees to the end-user. EOS^2 is able to replay arbitrarily structured multi-tier PHP applications with interleaved accesses to shared data. Our solution offers greatly enhanced programming convenience and thus productivity, with acceptable run-time overhead.

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, August 2004.

- [2] G. Shegalov and G. Weikum. EOS^2: Unstoppable stateful PHP (demo). 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. 1223–1226. ACM.

- [3] G. Shegalov, G. Weikum, and K. Berberich. Unstoppable stateful PHP web services. In K. Aberer, Z. Peng, E. A. Rundensteiner, Y. Zhang, and X. Li, eds., *Web Information Systems - WISE 2006, 7th International Conference on Web Information Systems Engineering*, Wuhan, China, 2006, *LNCS 4255*, pp. 132–143. Springer.

16.6 Query Processing and Optimization

Coordinator: Thomas Neumann

16.6.1 Efficient Top-k Evaluation

Investigators: Ralf Schenkel, Martin Theobald, and Gerhard Weikum in collaboration with Holger Bast and Debapriyo Majumdar

Top-k query processing is an important building block for ranked retrieval, with applications ranging from text and data integration to distributed aggregation of network logs and sensor data. Top-k queries operate on index lists for a query's elementary conditions and aggregate scores for result candidates. One of the best implementation methods in this setting is the family of threshold algorithms, which aim to terminate the index scans as early as possible based on lower and upper bounds for the final scores of result candidates. This procedure performs sequential disk accesses for sorted index scans, but also has the option of performing random accesses to resolve score uncertainty. This entails scheduling for the two kinds of accesses: 1) the prioritization of different index lists in the sequential accesses, and 2) the decision on when to perform random accesses and for which candidates.

The prior literature has studied some of these scheduling issues, but only for each of the two access types in isolation. Our IO-Top-k framework [1, 1] takes an integrated view of the scheduling issues and develops novel strategies that outperform prior proposals by a large margin. Our main contributions are new, principled, scheduling methods based on a Knapsack-related optimization for sequential accesses and a cost model for random accesses. The methods can be further boosted by harnessing probabilistic estimators for scores, selectivities, and index list correlations. They seamlessly integrate with our methods for probabilistic pruning that we developed earlier [4].

In performance experiments with three different datasets (TREC Terabyte, HTTP server logs, and IMDB), our methods achieved significant performance gains compared to the best previously known methods, namely different variants of Fagin's Threshold Algorithm [3]: a factor of up to 3 in terms of execution costs, and a factor of 5 in terms of absolute run-times of our implementation. Our best techniques are close to a lower bound for the execution cost of the considered class of threshold algorithms.

References

- [1] 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.
- [2] 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.

[3] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *J. Comput. Syst. Sci.*, 66(4):614–656, 2003.

[4] M. Theobald, G. Weikum, and R. Schenkel. Top-k query evaluation with probabilistic guarantees. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, eds., *Proceedings 2004 VLDB Conference, The 30th International Conference on Very Large Databases (VLDB)*, Toronto, Canada, 2004, pp. 648–659. Morgan Kaufmann. Acceptance ratio 1:6.

16.6.2 Distributed Top-k

Investigators: Sebastian Michel and Thomas Neumann in collaboration with Peter Triantafillou

Distributed top-k query processing has recently become an essential functionality in a large number of emerging application classes like Internet traffic monitoring and Peer-to-Peer Web search. We have developed efficient algorithms for distributed top-k queries in wide-area networks where the index lists for the attribute values (or text terms) of a query are distributed across a number of data peers. In [2], we have introduced KLEE, a powerful algorithm for distributed top-k query processing. KLEE is an extension of TPUT [1]. Consider a query that consists of m terms $t_1, ..., t_m$ with m peers $P_1, ..., P_m$ maintaining the index lists for the terms. In the first phase, the query initiator P_init fetches the top k entries from each of the involved index lists (peers) and aggregates the scores. We consider min_k to be the document currently at rank k. In phase 2, P_{init} sends the min_k/m threshold to the peers $P_1, ..., P_m$ that send back all $(docId, score)$-pairs with $score \geq min_k/m$. This ensures that all potential top-k candidate items have been observed at least one. Phase 3 is considered to be an optional phase where P_{init} retrieves the missing scores for all candidate documents using random lookups.

Considering TPUT [1] and our own algorithm KLEE [2], we have developed three different optimization techniques that use a sophisticated cost model to reason about parameter values. Our proposed approach uses Poison mixture models, or alternatively spline histograms to model the underlying score distribution. This knowledge is considered to be given a-priori.

- First, we introduce a technique to efficiently leverage the knowledge of the input data characteristics to tune score thresholds that are of fundamental importance. As for TPUT and KLEE, the threshold that is used in the second phase is uniform, i.e. P_{init} retrieves all index list entries that have a score greater than min_k/m. We propose the usage of non-uniform thresholds, i.e. we adapt the threshold to the index-lists score

distribution characteristics. The goal is to retrieve a minimum number of documents using m thresholds $\tau_1, ..., \tau_m$ with the constraint $\sum_i \tau_i = min_k$. As we address applications with large m the exact (optimal) solution is out of the question since it is hard to compute. However, we have developed an algorithm that computes an approximative solution that works pretty fine in practice.

– Second, we show how hierarchical query plans can be generated [3] using the aforementioned cost model to build optimal query execution plans that drastically increase the overall performance. Consider, for example, a query with one very large and several small input lists residing on different peers. It would be better to perform the top-k query at the peer with the large list, have the small peers ship their items to the large peer, and only send the final result to the query initiator. We use a dynamic programming approach that considers all possible query plans and chooses the cheapest plan w.r.t. our cost model.

– Third, we introduce a sampling method to select a subset of input data that still provides reasonably accurate results but can be, at the same time, more efficiently handled.

We have conducted a comprehensive performance evaluation using real-world data such as HTTP request logs and search engine query logs. All the aforementioned techniques result in a remarkably increased overall performance.

References

[1] P. Cao and Z. Wang. Efficient top-k query calculation in distributed networks. In S. Chaudhuri and S. Kutten, eds., *PODC*, 2004, pp. 206–215. ACM.

• [2] S. Michel, P. Triantafillou, and G. Weikum. KLEE: A framework for distributed top-k query algorithms. 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. 637–648. ACM. Acceptance ratio 1:6.

• [3] 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.

16.6.3 Join Order Optimization

Investigator: Thomas Neumann in collaboration with Guido Moerkotte

For the overall performance of a database management system, the cost-based query optimizer is an essential piece of software. One important and complex problem any cost-based query optimizer has to solve is that of finding the optimal join order. Existing database systems typically determine the optimal join order using dynamic programming strategies. Two principle strategies are commonly described in the literature: Either the the search space is organized by the size of the problem, i.e. the optimal join order is determined for an increasing number of relations, or it is organized by subsets, i.e. a join ordering problem is

solved by considering (already solved) subsets of the involved relations. The first strategy is popular in commercial database systems, while the second strategy is known to be fast for some specialized join ordering problems.

We analytically analyzed the time complexity of both algorithms for different kinds of join ordering problems, that is different kinds of query graphs [1]. The complexities and corresponding experiments showed two results: First, the two strategies different vastly depending on the considered class of join problems. The organization by size is superior for sparse search spaces, while the organization by subset is superior for dense search spaces. Second, both algorithms perform significantly worse than the known lower bound for join ordering dynamic programming strategies published in [3].

This motivated us to construct an algorithm that performs well for all kinds of join problems and in fact meets the lower bound for all dynamic programming join ordering strategies [1]. The basic idea is to reformulate the join ordering problem as a graph theoretic problem on the query graph: To be joined, two sub-problems must be connected by a join condition, i.e. an edge in the query graph. Thus, the problem of considering all possible join combinations can be reformulates as finding all connect pairs of connected subgraphs of the query graph. These pairs must be enumerated very efficiently (to meet the known lower complexity bound) and in an order suitable for dynamic programming.

We presented a suitable algorithm, and formally proved both its correctness and the achieved runtime complexity. The complexity meets the known lower bound not only asymptotically but exact, that is the algorithm considers the minimal number of join ordering problems. The experimental results have shown the algorithm is highly superior to the previously known strategies and indeed performs well for any kind of join ordering problem.

In addition, we examined the interaction of maps (i.e. computed values), selections and joins under factorization [2]. When factorizing common expressions, the costs of selections and joins affect each other, as one operator might produce computed values required by a following one. We had shown in a previous paper that the problem is NP hard in general when only considering maps and selections, and presented suitable algorithms. In [2] we generalized them to joins, reducing the search space increase as much as possible. The experimental results confirmed that, first, factorization is a critical issue when it comes to generating optimal plans and second, considering factorization does not make plan generation significantly more expensive for typical queries.

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, September 2006, pp. 930–941. ACM. Acceptance ratio 1:7.

- [2] T. Neumann, S. Helmer, and G. Moerkotte. On the optimal ordering of maps, selections, and joins under factorization. In D. J. Bell and J. Hong, eds., *Flexible and efficient information handling, 23rd British National Conference on Databases, BNCOD 23*, Belfast,Northern Ireland, UK, September 2006, *LNCS 4042*, pp. 115–126. Springer.

[3] K. Ono and G. M. Lohman. Measuring the complexity of join enumeration in query optimiza-

tion. In D. McLeod, R. Sacks-Davis, and H.-J. Schek, eds., *VLDB*, 1990, pp. 314–325. Morgan Kaufmann.

16.6.4 Time Travel Queries

Investigators: Klaus Berberich, Srikanta Bedathur, and Thomas Neumann

A number of large-scale evolving text collections are available today. Web archives that are the outcome of efforts such as Internet Archive [5] and collaboratively edited Wikipedia encyclopedia [1] are the prime examples of such evolving text collections. In addition to their content, these collections hold even richer latent knowledge in the form their *evolutionary history* of content and structure. It not only captures the evolution of digital content but also embodies the near-term history of our society, economy, and science. Despite the ubiquity of these data sources, there is a surprising lack of satisfactory tools that can mine and extract the valuable knowledge from them.

Consider an illustrative scenario of a case of copyright-protection where a lawyer needs to collect evidence about the illegal availability of copyrighted material (e.g., a piece of writing) on the Web. A keyword query to a Web search-engine may not show up any results if the material is no longer available on the Web. Although contents of the Internet Archive may contain such an evidence, the only form of exploring archived contents currently available is through a system such as the Wayback Engine, which requires one to specify the exact URL whose evolutionary history is sought. It is very unlikely that the investigating lawyer is aware of the exact URL from where evidence has to be collected. Similar scenarios abound in many other contexts such as the investigative journalism, market research on the evolution of product reviews etc., where historical information is sought.

The main reason why these rich sources of information are underexploited is due to the problem of their scale. Collections like the Web or Wikipedia are massive even if only their recent snapshot is considered, requiring massive storage, computing power, and efficient algorithms for searching. Looking at their complete evolutionary history we are faced with even larger and ever growing data volumes, hence the challenge is even greater. As a consequence, novel approaches to text search over temporally evolving collections that provide excellent scalability are needed.

In this work, we address this challenge by introducing the notion of *time-travel queries* over these versioned collections, and developing compact synopses structures as well as efficient query evaluation strategies. Time-travel queries are aimed at supporting the evolutionary (temporal) analysis over Web archives extending the power of Web search-engines. 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. It is required that the Q_{ir} be evaluated and ranked based on the content and structure of the archived collection at the historical time period defined by Q_{tc}.

In [2], we developed a scalable framework for indexing large-scale evolving text collections and efficiently answering time-travel queries over them. First of all, we proposed a relevance model for time-travel queries based on state-of-art Okapi-BM25 [7, 8] probabilistic content relevance model, and generate document-level rankings by aggregating version-level scores across the specified temporal context Q_{tc}. Next, we developed an indexing infrastructure that enriches the popular inverted file index with temporal information of versions. Our

infrastructure overcomes the resulting index-size blow up by exploiting the fact that most inter-version changes of a document are not significant enough to change its rank in query results. Finally, we extended no-random-access (NRA) variant of the Threshold Algorithm [4] family for generating top-k results, to work with time-travel queries. Our initial experiments using the revision history of Wikipedia [1] showed the effectiveness of our proposal.

In order to obtain effective results for time-travel queries, just as in the case of Web search, it is expected that the answer consist of a list of documents that are ranked based on both the content-wise relevance with the query terms, as well as a query-independent measure that reflects *authority/importance/prestige* for each document. Motivated by this, in [3], we developed a compact synopses structure called *Rank Synopses* using which it is possible to accurately reconstruct historical PageRank scores [6], a prevalent authority measure for the Web. Rank synopses views normalized PageRank values (refer Section 16.7.1) across time as a time-series, and applies techniques from time-series research to compactly store them.

Our current research is geared towards integration of rank synopses with the indexing infrastructure for time-travel queries, and significant performance enhancements to the indexing infrastructure itself.

References

[1] http://www.wikipedia.org.

- [2] 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.

- [3] K. Berberich, S. J. Bedathur, and G. Weikum. Rank synopses for efficient time travel on the web graph. 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. 864–865. ACM.

[4] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *J. Comput. Syst. Sci.*, 66(4):614–656, 2003.

[5] http://www.archive.org.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[7] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1994, pp. 232–241. Springer-Verlag New York, Inc.

[8] S. E. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In *TREC*, 1999.

16.7 Web and Text Mining

Coordinator: Srikanta Bedathur

16.7.1 Temporal Link Analysis

Investigators: Klaus Berberich and Srikanta Bedathur in collaboration with Michalis Vazirgiannis

Link analysis techniques are applied by web search engines to assess the authority of web pages. This assessment of authority is then used (together with other factors) to rank the web pages that are textually relevant to a given query. Link analysis techniques are based on the idea that hyperlinks between web pages can be interpreted as recommendations and therefore exploit the Web's link structure to assess authority. The HITS method proposed by Kleinberg [6] and the PageRank method proposed by Page et al. [9] are the seminal papers in this field; a survey of the ample work spanned by these early papers is given by [7].

All prior methods are applied to a static, possibly partial snapshot of the web graph and hence disregard the high dynamics of the web graph [8]. In contrast, our work considers the web graph's dynamics and defines two new classes of link analysis techniques, namely *time-aware* and *trend-based link analysis*.

Time-aware link analysis takes into account the temporal dimension of the user's interest and produces rankings reflecting it. For many queries users have a precise idea of the temporal dimension of their interest. For a query relating to the Olympics opening ceremony, as an example, a user could be interested in the 2004 summer Olympics that took place in Athens or the more recent 2006 winter Olympics in Torino.

The *T-Rank* family of techniques proposed in [3, 4] therefore explicitly considers a user-specified time interval of interest – the user's temporal interest. Based on the temporal interest two temporal features of web pages and hyperlinks resulting from the Web's dynamics, namely *freshness* and *activity* are computed. The *freshness* of a web page or a hyperlink, on the one hand, conveys how recent it is with regard to the user's temporal interest. The *activity* of a web page or a hyperlink, on the other hand, expresses the frequency of change with regard to the user's temporal interest. Freshness and activity are then used by the *T-Rank* family of techniques to redefine PageRank's random walk on the web graph. To this end, freshness and activity are combined in a weighted manner with weighting coefficients opening further opportunities to adapt to the user's needs. The effectiveness of the *T-Rank* family of ranking methods was shown through user studies in a series of experiments across different datasets.

Trend-based link analysis, on the other hand, produces rankings that reflect which pages were gaining relatively most authority in a specified time interval of interest and that therefore reflect the Zeitgeist of that epoch. For a research-related query, for instance, trend-based link analysis ranks publications in currently hot topics above seminal publications that have stood the test of time.

The *BuzzRank* method proposed in [2] analyzes time series of normalized PageRank scores to quantify the relative authority change of a web page in a given time interval of interest. As an enabling technique for *BuzzRank* we developed a normalization scheme for PageRank scores [1] making them comparable across different graphs. *BuzzRank* builds on the observation made by Cho et al. [5] that the authority evolution of a web page can be described by a

logistic growth model. Since the time interval of interest is usually short, *BuzzRank* assumes that the authority growth rate during the time interval is fixed, thus yielding a simpler exponential growth model. The two parameters of this exponential growth model, conveying the absolute level of authority and its growth rate, are then estimated using the observations of the time series during the time interval of interest. Finally, *BuzzRank* obtains a trend-based ranking based on the estimated authority growth rate of web pages. Experiments on an evolving graph derived from the Digital Bibliography and Library Project (DBLP) gave promising anecdotic evidence for the effectiveness of *BuzzRank*.

One topic in our ongoing work is the integration of the *T-Rank* family of methods and the *BuzzRank* method with our recently developed time-travel search techniques (see Section 16.6.4).

References

- [1] K. Berberich, S. Bedathur, M. Vazirgiannis, and G. Weikum. Comparing apples and oranges: Normalized pagerank for evolving graphs. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, May 2007. ACM.
- [2] K. Berberich, S. J. Bedathur, M. Vazirgiannis, and G. Weikum. Buzzrank ... and the trend is your friend. In *Proceedings of the 15th International World Wide Web Conference (WWW 2006)*, Edinburgh, UK, 2006, pp. 937–938. ACM.

[3] K. Berberich, M. Vazirgiannis, and G. Weikum. T-rank: Time-aware authority ranking. In S. Leonardi, ed., *Algorithms and Models for the Web-graph, Third International Workshop, WAW 2004*, Rome, Italy, October 2004, *LNCS 3243*, pp. 131–142. Springer.

- [4] K. Berberich, M. Vazirgiannis, and G. Weikum. Time-aware authority ranking. *Internet Mathematics*, 2(3):301–332, 2006.

[5] J. Cho, S. Roy, and R. E. Adams. Page Quality: in Search of an Unbiased Web Ranking. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of Data*, New York, NY, USA, 2005, pp. 551–562. ACM Press.

[6] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[7] A. N. Langville and C. Meyer. Deeper Inside PageRank. *Internet Mathematics*, 1(3):335–380, 2004.

[8] A. Ntoulas, J. Cho, and C. Olston. What's New on the Web?: The Evolution of the Web from a Search Engine Perspective. In *Proceedings of the 13th conference on World Wide Web*, 2004, pp. 1–12. ACM Press.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

16.7.2 User Behavior

Investigators: Julia Luxenburger and Gerhard Weikum

The analysis of observed user search and browsing behavior is a valuable information source for many aspects of web search result ranking. Even in the face of noise, it offers better

grounds for relevance feedback than mere pseudo-relevance feedback would give which simply treats the top-k retrieved documents as relevant.

From the monitoring of user interactions with a search engine we are able to draw conclusions of different flavor:

- The sequence of queries, a user subsequently poses, allows us to group related queries serving the same information need and learn query reformulation patterns.
- The query-result pages which were clicked on and the ones which were not clicked on after a user saw the summary snippets of the top-10 results, lead us to inferences on the relevance, respectively irrelevance, of result pages to their corresponding queries, as well as to inferences on the general quality of these pages.
- The analysis of complete user search sessions enables us to identify frequent user interaction patterns, as well as deficiencies of state-of-the-art web search [1].

Such information can be employed in the course of various facets of search result ranking, both in the amelioration of single ranking features, such as the similarity between document and query based on their textual content, as well as in determining a meaningful combination of these ranking features to form the final ranking. Furthermore the granularity of the gathered user monitoring data, whether a single user or a larger community of users is considered, calls for a different usage of the available data, e.g., targeting at a personalization of search results as opposed to make users benefit from insights derived from the whole user base.

One focus of our work in this area is the incorporation of implicit user feedback into Web link analysis which constitutes an important ranking feature. State-of-the-art authority analysis methods on the Web linkage graph such as the PageRank [3] algorithm are based on the assumption that a web page author endorses a Web page when creating a hyperlink to that page. This kind of intellectual user input can be generalized to a user endorsing a query-result page when visiting that page, and moreover disapproving a result page when preferring a lower-ranked result page.

We study link analysis methods [2] that enhance PageRank by incorporating additional user assessments based on query logs and click streams, including negative feedback when a query-result page does not satisfy the user demand or is even perceived as spam. Our methods use various novel forms of Markov reward models whose states correspond to users and queries in addition to Web pages and whose links also reflect the relationships derived from query-result clicks, query refinements, and explicit ratings.

Our experiments, based on real-life query-log and click-stream traces on an excerpt of the English version of the Wikipedia encyclopedia, indicate the potential of our methods.

References

• [1] N. Kammenhuber, J. Luxenburger, A. Feldmann, and G. Weikum. Web search clickstreams. In J. M. Almeida, V. A. F. Almeida, and P. Barford, eds., *Proceedings of the 6th ACM SIGCOMM on Internet measurement (IMC '06)*, Rio de Janeiro, Brazil, 2006, pp. 245–250. ACM.

- [2] J. Luxenburger and G. Weikum. Exploiting community behavior for enhanced link analysis and web search. In D. Zhou, ed., *Proceedings of the 9th International Workshop on the Web and Databases (WebDB 2006)*, Chicago, Illinois, USA, 2006, pp. 14–19.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

16.7.3 Ontology Creation

Investigators: Gjergji Kasneci and Fabian Suchanek

Many applications in modern information technology rely on ontological background knowledge. This applies foremost to applications in the scope of the Semantic Web, but also to Document Classification, Machine Translation, Word Sense Disambiguation, search, and many more. These applications could boost their performance if a huge ontology with knowledge from several sources was available. With our ontology YAGO, this goal has come within reach (see [4]).

YAGO is a light-weight and extensible ontology with high coverage and quality. YAGO builds on entities and relations and currently contains more than 1 million entities and roughly 5 million facts about them. This includes the Is-A hierarchy as well as non-taxonomic relations between entities (such as e.g. `hasWonPrize`). The facts have been extracted automatically from Wikipedia and unified with WordNet, using a carefully designed combination of rule-based and heuristic methods. Our empirical evaluation of fact correctness shows an accuracy of about 95%. YAGO is based on a logically clean model, which is decidable, extensible, and compatible with RDFS.

YAGO's data model supports reification (i.e., it assigns identifiers to facts). Thereby the model can express relations between facts (e.g., popularity rankings of pairs of soccer players and their teams) and general properties of relations (e.g., transitivity or acyclicity). Still, YAGO's data model is decidable – in contrast to OWL Full, which is by now the only version of the Web Ontology Language (OWL) that can also express relation properties.

YAGO's core is assembled from one of the most comprehensive lexicons available today, Wikipedia. Our Information Extraction approach makes use of the Wikipedia category pages. Category pages are lists of articles that belong to a specific category (e.g., Zidane is in the category of *French football players*). These lists give us candidates for entities (e.g. Zidane), candidates for concepts (e.g. IsA(Zidane, FootballPlayer)) and candidates for relations (e.g. isCitizenOf(Zidane, France)). In an ontology, concepts have to be arranged in a taxonomy to be of use. The Wikipedia categories are indeed arranged in a hierarchy, but this hierarchy is barely useful for ontological purposes. For example, Zidane is in the super-category *Football in France*, but Zidane is a football *player* and not a football. WordNet, in contrast, provides a clean and carefully assembled hierarchy of thousands of concepts. But the Wikipedia concepts have no obvious counterparts in WordNet. To the best of our knowledge, our method is the first approach that accomplishes this unification between WordNet and facts derived from Wikipedia with an accuracy of 97%. In order to map a Wikipedia category name to a WordNet synset, we proceed as follows: We first determine the head compound, the premodifier and the post-modifier of the category name. For example, for the Wikipedia category *American people in Japan*, these are "American", "people" and "in Japan", respectively. We stem the head compound of the category name (i.e. *people*) to its singular form (i.e. *person*).

Then we check whether there is a WordNet synset for the concatenation of pre-modifier and head compound (i.e. *American person*). If this is the case, the Wikipedia category becomes a subconcept of the WordNet concept. If this is not the case, we exploit the fact that the Wikipedia category names are almost exclusively endocentric compound words (i.e. the category name is a hyponym of its head compound, e.g. "American person" is a hyponym of "person"). The head compound (*person*) has to be mapped to a corresponding WordNet synset. After several experiments, we found out that mapping the head compound simply to the most frequent synset[4] yields the correct synset in the overwhelming majority of cases. This way, the Wikipedia class *American people in Japan* becomes a subclass of the WordNet class *person/human*. This methodology also proved to be useful in the construction of a large semantically annotated XML corpus described in [1].

We exploit other Wikipedia categories to extract relationship instances for the relations BORNINYEAR, DIEDINYEAR, ESTABLISHEDINYEAR, LOCATEDIN, WRITTENINYEAR, POLITICIANOF, and HASWONPRIZE. Furthermore, we utilize Wikipedia's redirect system to find alternative names for entities (e.g. redirect from "Einstein, Albert" to the page "Albert Einstein" indicates that "Einstein, Albert" is an alternative name for the entity *Albert Einstein*). We depict a fragment of YAGO in the following Figure.

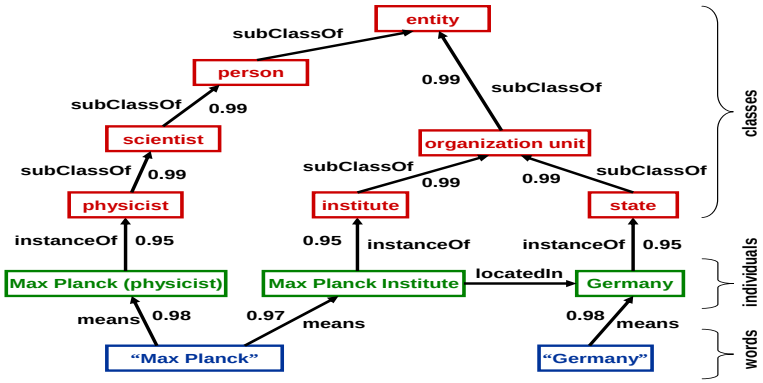


Figure 16.4: An example fragment of the YAGO ontology

YAGO is designed to be easily extendable. We use our system LEILA[5], which is described in [2], for this purpose. LEILA can extract semantic information from natural language texts. Given a semantic relation (such as e.g. `bornInYear`) and a natural language corpus, LEILA can extract pairs of entities that stand in the given relation (e.g. `Einstein/1879`). LEILA combines linguistic analysis and machine learning techniques to find robust patterns in the text and to generalize them.

Different from previous approaches, LEILA uses *syntactic* patterns instead of surface patterns, i.e. it stores the grammatical relations between the pattern constituents. Thereby the patterns are more robust to variations in the sentences. Furthermore, LEILA collects not only positive patterns, but also negative patterns. It trains a classifier on these patterns and thereby prevents the over-generalization of patterns. We have compared LEILA to various

[4] WordNet assigns to each word the frequencies with which it refers to possible synsets.
[5] Learning to Extract Information by Linguistic Analysis

state-of-the-art competitors and LEILA shows a superior performance.

References

- [1] R. Schenkel, F. M. 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.
- [2] F. M. 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.
- [3] F. M. Suchanek, G. Ifrim, and G. Weikum. LEILA: Learning to extract information by linguistic analysis. In P. Buitelaar, P. Cimiano, and B. Loos, eds., *Proceedings of the 2nd Workshop on Ontology Learning and Population (OLP2) @COLING/ACL 2006*, Sydney, Australia, 2006, pp. 18–25. ACL. Acceptance Rate: 8/19=42
- [4] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge - unifying WordNet and Wikipedia. In *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007. IW3C2.

16.7.4 Meta Classification and Clustering

Investigator: Stefan Siersdorfer

Automatic document classification and clustering are useful for a wide range of applications such as organizing Web, intranet, or portal pages into topic directories, filtering news feeds or mail, focused crawling on the Web or in intranets, and many more.

We have developed a family of adjustable metamethods based on ensemble learning that can be tuned towards the specific goals of the classifier: when precision and accuracy are critical we can prioritize the mutual agreement of multiple classifiers, but when recall is more important we can relax the settings of the meta-classifier. A particular application of these methods is our focused crawler BINGO!, which can start with a small set of training samples and seed URLs and then automatically finds characteristic "archetype" pages for dynamic and automated re-training. In this semisupervised-learning context, it is crucial to minimize the classification error, and restrictive metamethods can be effectively tuned towards this goal.

Our recent work in this direction [1] addresses the problem of semisupervised classification on document collections using retraining (also called self-training). In the context of our BINGO! focused crawler, the key challenge in reducing the classification error is the selection of parameters such as the number of selected documents, the number of retraining iterations, and the ratio of positive and negative classified used for retraining etc. Our work develops methods for automatically tuning these parameters based on predicting the leave-one-out error for a retrained classifier and avoiding the dilution of classifier from selection of too many weak documents for retraining. Our experiments with a variety of datasets confirm the practical viability of our solutions.

References

- [1] S. Siersdorfer and G. Weikum. Automated retraining methods for document classification and their parameter tuning. In A. H. H. Ngu, M. Kitsuregawa, E. J. Neuhold, J.-Y. Chung, and Q. Z. Sheng, eds., *Web information systems engineering - WISE 2005, 6th International Conference on Web Information Systems Engineering*, New York, USA, 2005, *LNCS 3806*, pp. 478–486. Springer.

16.7.5 Graph-based Classification and Clustering

Investigators: Ralitsa Angelova and Stefan Siersdorfer

Suppose we are asked to divide companies selling products on eBay into two sets, roughly representing trustworthy vs. untrustworthy ones. It would be of great help if we could access all forums and customer pages that discuss the products and companies in question and their quality. This helpful information can be revealed only if we follow the existing links among the data items in the corpus. In other words, link information provides refined clues about relationships in the data set that can later crucially influence the final data partitioning. Nevertheless, content-based partitioning approaches are completely ignorant of it. Graph-based methods, on the other hand, consider the hyperlinked corpora without omitting the valuable link structure. Typically, the underlying graph G is constructed by nodes, representing data points, and edges, representing relationship among the connected nodes. Each edge carries a weight, indicating the strength of the relationship, usually derived from the similarity of the connected node pair.

To automatically structure heterogeneous document collections into thematically coherent subsets we need to find the right partitioning of the data. The issue is relevant for a variety of applications, such as structuring large amounts of enterprise and intranet data, dividing topics in large web directories into subtopics, organizing large personal email folders, etc. This problem can be viewed as a supervised classification problem if explicit manually labeled training data is available. In a traditional classification problem, we wish to assign one of k labels (or classes) to each of n objects (or documents) in a way that is consistent with some observed data (training data) available about that problem.

Aware of the predictive power of the link structure, we model the mutual influence between neighboring documents in the graph to estimate the class labels of all test documents simultaneously. Our goal is to find the maximally likely labeling of a given graph G that minimizes the sum of two types of costs: *assignment* and *separation* cost [7]. The first cost is based on the individual choice of a label which is assigned to a document. The latter reflects the choice of a pair of labels to which two neighboring documents belong. The problem has been shown to be NP-complete [7]. In principle, the graph model could consider long-range influences among transitively related documents, with decreasing influence as the distance in the graph increases. For tractability, however, it makes sense to focus on the strongest dependencies among immediate neighbors. Such a model is called a first-order Markov Random Field or MRF [8, 9]. Computing the parameters of an MRF, such that the likelihood of the observed labels is maximized, is a difficult problem that cannot be solved in closed analytic form and is typically addressed by an iteration technique known as relaxation labeling (RL) [3, 4]. Our approach builds on this mathematical technique. Instead of seeking a global optimization, which obtains a unique labeling for all nodes in the test graph, it starts with a greedy labeling of the graph, paying attention only to the node-labeling (assignment) cost,

and then iteratively "corrects" the neighborhood labeling where the presence of edges leads to a very high penalty in terms of the separation costs. In contrast to the prior work on this problem we propose to capture and exploit the link/class patterns in the complete graph *dynamically* as RL iterates. For enhanced robustness, we select a "reliable" set of neighbors for each test document and discriminate different levels of trust in different neighbors by assigning *weights* to links based on the content similarity of nodes. Lead by the intuition that neighboring documents should receive similar class labels, we propose to use a *distance metric* over the set of possible labels that helps the classifier to distinguish thematically closely related labels from those that are thematically too far apart. This is crucial when the training data size is small.

However, explicit training data is always very expensive to gather and often even unavailable. Therefore, the only viable option might be to view our structuring problem as an unsupervised *clustering* problem. That is, we aim to partition or cluster the data according to the clustering hypothesis, which states that closely associated documents tend to be relevant to the same requests, hence, located in one cluster.

We propose a powerful and flexible framework [1, 2], abbreviated NECKLACE, which can benefit from the link structure and exploit the predictions of any clustering algorithm: graph or content based. NECKLACE (NEighborhood-conscious Clustering Kit with Link Awareness for Cluster Enhancement) is a probabilistic method inspired by the novel classification method presented above, but applied in the (unsupervised) graph-clustering scenario where no prior knowledge about the data is available to the algorithm. A salient feature of our algorithm is its applicability as a last-stage refinement of the cluster purity produced by any initial clustering algorithm: content- or graph-based. The extensive experimental study on three real world data sets (DBLP, IMDB, and Wikipedia) shows the superiority of NECKLACE in comparison to content-based (e.g., traditional [5] or bisection-based [6] k-Means), hybrid, and graph-cut based clustering methods [6]. The achieved absolute gains in clustering accuracy of 10%, 5% and 8%, respectively, are highly significant.

References

- [1] R. Angelova and S. Siersdorfer. A neighborhood-based approach for clustering of linked document collections. In P. S. Yu, V. J. Tsotras, E. A. Fox, and B. Liu, eds., *Proceedings of the Conference on Information and Knowledge Management, CIKM 2006*, Arlington, VA, USA, 2006, pp. 778–779. ACM.

- [2] R. Angelova and S. Siersdorfer. A neighborhood-based approach for clustering of linked document collections. Research Report MPI-I-2006-5-005, Max-Planck-Institut for Informatics, Saarbrücken, Germany, October 2006.

- [3] R. Angelova and G. Weikum. Graph-based text classification: Learn from your neighbors. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Jaervelin, eds., *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, 2006, pp. 485–492. ACM. Acceptance ratio 1:5.

[4] S. Chakrabarti, B. E. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, 1998, pp. 307–318.

[5] J. Hartigan and M. Wong. A k-means clustering algorithm. *Applied Statistics, 28:100-108*, 1979.

[6] G. Karypis. Cluto: A software package for clustering high-dimensional data sets. (www-users.cs.umn.edu/karypis/cluto), 2003.

[7] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. In *FOCS*, 1999, p. 14.

[8] S. Z. Li. *Markov random field modeling in image analysis.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[9] L. Pelkowitz. A continuous relaxation labeling algorithm for markov random fields. *IEEE Transactions on Systems, Man, and Cybernetics*, 20:709–715, 1990.

16.7.6 Transductive Classifiers

Investigators: Georgiana Ifrim and Gerhard Weikum

Many applications require classification models able to learn from few labeled data and rich background knowledge. Examples could be Amazon organizing book descriptions into predefined categories, Google or Yahoo! classifying crawled Web pages into topic directories for more convenient access, or Wikipedia categorizing encyclopedia articles for better search and browsing. A learning paradigm in which the data collection to be automatically labeled is available beforehand is referred to as *transductive inference*.

Transductive learning is particularly attractive for text classification with very few explicitly labeled training documents, which happens whenever human assessment is the (time or cost) bottleneck on rapidly growing and highly diverse corpora. In such a setting, being able to harness the feature distributions and relations among the unlabeled documents is an important asset to improve the classifier. Another potentially beneficial asset is explicit knowledge about concepts, words and phrases that express concepts, and the semantic relations among concepts (e.g., hyponymy). Such knowledge sources may be given in the form of an ontology or thesaurus.

Prior work has mostly pursued "latent semantic" models such as spectral analysis, and the few approaches that have attempted to leverage explicit knowledge sources focused on concept-aware feature spaces and did not integrate concept-word relationships into the learning procedure itself. Moreover, many of these prior methods faced difficult model selection problems regarding feature engineering and parameter tuning.

We propose a novel approach, based on a generative model with explicit concepts, that combines background ontologies with transductive learning on large corpora [3]. Our techniques aim at making classifiers more robust when very few training data is available, by exploring existing background knowledge sources.

Our approach is based on a generative model for text documents where words are generated by concepts which in turn are generated by topics. We postulate conditional independence between words and topics given the concepts. Once the corresponding probabilities for word-concept and concept-topic pairs are estimated, we can use Bayesian inference to compute the probability that a test document with known words but unobservable concepts belongs to a certain topic. The concepts are used as latent variables, but unlike earlier work on spectral decomposition and latent semantic models [1], [2] our concepts are named and can be explicitly identified in the underlying ontology or thesaurus. We employ an iterative EM (Expectation-Maximization) procedure for maximum-likelihood parameter estimation. The

effectiveness of EM greatly benefits from a judicious initialization step that estimates word-concept probabilities on the *unlabeled* part of the document collection, thus leading us to a transductive learning method.

Our contributions can be summarized as follows:

1. By using explicit concepts from an ontology or thesaurus and by using a heuristic technique for bootstrapping the word-to-concept mapping, we avoid the model selection problem inevitably faced by all techniques based on latent dimensions (i.e., choosing an appropriate number of dimensions).

2. By the same token, we avoid the combinatorial explosion in the space of parameters to be estimated (i.e., concept-word pairs), and we largely eliminate the need for parameter smoothing (which is often a very tricky issue).

3. The initial word-to-concept mapping is very beneficial for fast convergence of EM, and reduces the danger of getting stuck in local maxima of the likelihood function. The latter is an issue especially with very few training data.

4. Our method provides an intuitive and effective way of exploiting the available resources, unlabeled documents from the full corpus and ontological relationships among concepts, resulting in improved classification accuracy with few training data.

5. Our approach is more robust in the sense that it requires considerably less tuning than other transductive methods.

In our experiments, with real-life datasets from the Reuters newswire corpus, Amazon book reviews, and Wikipedia articles, we compare our method with the Spectral Graph Transducer [5] and Transductive SVM classifiers [4] and demonstrate the viability and superiority of our approach.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. In *NIPS*, 2002.

[2] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1), 2001.

- [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] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, 1999.

[5] T. Joachims. Transductive learning via spectral graph partitioning. In *ICML*, 2003.

16.7.7 Multilingual Classifiers

Investigators: Gerard de Melo and Stefan Siersdorfer

While there has been a considerable amount of research on automatically classifying documents from monolingual text collections into different categories, the same cannot be said of the case of collections with documents given in multiple distinct languages. For such scenarios, we have developed a method of representing text as feature vectors that can be used with machine learning algorithms in order to learn classifications from pre-classified examples and then automatically classify documents that might be provided in completely different languages [1]. Our approach, called Ontology Region Mapping, uses either multiple aligned WordNet-like resources or machine translation combined with a monolingual resource such as Princeton WordNet. In order to overcome lexical differences between languages, it goes beyond a direct mapping from terms to concepts, instead mapping to entire regions of concepts and thus fully exploiting the external knowledge manifested in the resources. A graph traversal algorithm is used to determine which related concepts are most likely to be relevant and to what degree they should be considered. Extensive testing has shown that this method leads to significant improvements compared to existing approaches.

References

- [1] G. de Melo and S. Siersdorfer. Multilingual text classification using ontologies. In G. Amati, ed., *Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007)*, Rome, Italy, 2007, *LNCS 4425*. Springer. Acceptance Ratio 1:4.

16.8 Academic Activities

16.8.1 Journal Positions

Gerhard Weikum is on the editorial board of

- IEEE Computer Society Transactions on Knowledge and Data Engineering (TKDE)

16.8.2 Book Series Positions

Gerhard Weikum is on the editorial board of

- Springer-Verlag Series "Lecture Notes in Computer Science (LNCS)"
- Editorial Board, Springer-Verlag Series "Data-centric Systems and Applications (DCSA)"

16.8.3 Conference and Workshop Positions

Membership in Program Committees

Ralitsa Angelova:

- *2nd International Conference on Scalable Information Systems (INFOSCALE 2007), Suzhou, China, June 2007*

Sebastian Michel:

- *2nd International Conference on Scalable Information Systems (INFOSCALE 2007), Suzhou, China, June 2007*

Thomas Neumann:

- *27th International Conference on Distributed Computing Systems (ICDCS 07)*, Toronto, June 2007
- *11th East-European Conference on Advances in Databases and Information Systems (ADBIS 07)*, Varna, Bulgaria, September 2007
- *24th International Conference on Data Engineering (ICDE 2008)*, Cancun, Mexico, April 2008

Maya Ramanath:

- *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, May 2007

Ralf Schenkel:

- *14th ACM Conference on Information and Knowledge Management (CIKM 2005)*, Bremen, Germany, November 2005
- *2nd Workshop on Database Technologies for Handling XML Information on the Web (DataX 2006)*, Munich, Germany, March 2006
- *Workshop on Query Languages and Query Processing (QLQP 2006)*, Munich, Germany, March 2006
- *32nd International Conference on Very Large Databases (VLDB 2006)*, Seoul, Korea, September 2006
- *PhD Workshop at the 32nd International Conference on Very Large Databases (VLDB 2006)*, Seoul, Korea, September 2006
- *28th European Conference on Information Retrieval (ECIR 2006)*, London, United Kingdom, April 2006
- *German Computer Society Workshop on Information Retrieval*, Hildesheim, Germany, October 2006
- *12th Conference on Databases in Business, Technology, and the Web (BTW 2007)*, Aachen, Germany, March 2007
- *29th European Conference on Information Retrieval (ECIR 2007)*, Rome, Europe, April 2007
- *30th International ACM SIGIR Conference (SIGIR 2007)*, Amsterdam, The Netherlands, July 2007
- *18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, Regensburg, Germany, September 2007
- *33rd International Conference on Very Large Data Bases (VLDB 2007), Demo Track*, Vienna, Austria, September 2007

- *German Computer Society Workshop on Information Retrieval*, Halle, Germany, September 2007
- *15th International Conference on Cooperative Information Systems (CoopIS 2007)*, Iberian peninsula and islands, October 2007
- *11th International Conference on Extending Database Technology (EDBT 2008)*, Nantes, France, March 2008

Stefan Siersdorfer:

- *29th European Conference on Information Retrieval (ECIR 2007)*, Rome, Italy, April 2007

Martin Theobald:

- *9th International Workshop on the Web and Databases (WebDB 2006)*, Chicago, USA, June 2006
- *29th European Conference on Information Retrieval (ECIR 2007)*, Rome, Italy, April 2007

Christos Tryfonopoulos

- *3rd International Conference on Innovations in Information Technology (IIT 2006)*, Dubai, November 2006
- *4th International Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR 2006)*, Arlington, USA, November 2006
- *2nd International Conference on Scalable Information Systems (INFOSCALE 2007), Suzhou, China, June 2007*
- *11th East-European Conference on Advances in Databases and Information Systems (ADBIS 2007)*, Varna, Bulgaria, September 2007

Gerhard Weikum:

- *8th International Workshop on the Web and Databases (WebDB 2006)*, Baltimore, USA, June 2005
- *31st International Conference on Very Large Data Bases (VLDB)*, Trondheim, Norway, August 2005
- *9th European Conference on Digital Libraries (ECDL)*, Vienna, Austria, September 2005
- *2nd International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)*, Atlanta, USA, April 2006
- *15th International World Wide Web Conference (WWW 2006)*, Edinburgh, UK, May 2006
- *12th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, Philadelphia, USA, August 2006
- *2006 ACM SIGMOD International Conference on Management of Data (SIGMOD 2006)*, Chicago, USA, June 2006

- *9th International Workshop on the Web and Databases (WebDB 2006)*, Chicago, USA, June 2006
- *German Computer Society Workshop on Information Retrieval*, Hildesheim, Germany, October 2006
- *32nd International Conference on Very Large Databases (VLDB 2006)*, Seoul, Korea, September 2006
- *6th International Workshop on Web Archiving*, Alicante, Spain, September 2006
- *15th International Conference on Information and Knowledge Management (CIKM 2006)*, Arlington, USA, November 2006
- *International Workshop on Peer-to-Peer Information Retrieval (P2PIR)*, Arlington, USA, November 2006
- *3rd Biennial Conference on Innovation Data Systems Research (CIDR 2007)*, Asilomar, USA, January 2007
- *12th German Conference on Databases in Business, Technology, and the Web (BTW 2007)*, Aachen, Germany, March 2007
- *23rd IEEE International Conference on Data Engineering (ICDE 2007)*, Istanbul, Turkey, April 2007
- *International Workshop on Self-Managing Database Systems (SMDB)*, Istanbul, Turkey, April 2007
- *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, May 2007
- *2007 ACM SIGMOD International Conference on Management of Data (SIGMOD 2007)*, Beijing, China, June 2007
- *2007 ACM SIGIR International Conference on R&D in Information Retrieval (SIGIR 2007)*, Amsterdam, The Netherlands, August 2007
- *16th ACM Conference on Information and Knowledge Management (CIKM 2007)*, Lisboa, Portugal, November 2007
- *24th International Conference on Data Engineering (ICDE 2008)*, Cancun, Mexico, April 2008

Membership in Organizing Committees

Ralf Schenkel:

- *Relevance Feedback Task at INEX 2006* – co-organizer

Gerhard Weikum:

- *Dagstuhl Seminar 06121 on 'Atomicity: A Unifying Concept in Computer Science'*, Schloss Dagstuhl, April 2006 – Jo-organizer
- *3rd Biennial Conference on Innovative Data Systems Research (CIDR 2007)*, Asilomar, January 2007 – co-chair and PC chair

16.8.4 Invited Talks and Tutorials

Sebastian Michel:

- EPFL Lausanne, February 2007
- CWI Amsterdam, March 2007

Thomas Neumann:

- Mannheim University, February 2007

Ralf Schenkel:

- HPI Potsdam, July 2005
- TU Clausthal, February 2006
- Ulm University, December 2006
- Queen Mary University, London, March 2007

Gerhard Weikum:

- Keynote, 11th German Conference on Database Systems for Business, Technology, and Web (BTW), Karlsruhe, Germany, March 2005
- University of Waterloo, Canada, June 2005
- Tutorial at ACM SIGMOD Conference, Baltimore, USA, June 2005
- European Commission, IST Department on Future and Emerging Technologies (FET), Brussels, Belgium, December 2005
- Keynote, International Workshop on Algorithmic and Numerical Aspects in Web Search, Pisa, Italy, February 2006
- Tutorial at IEEE International Conference on Data Engineering (ICDE), Atlanta, USA, April 2006
- University of Toronto, Canada, April 2006
- AT&T Labs, Florham Park, USA, April 2006
- ASK/Teoma, Piscataway, USA, April 2006
- Keynote, International Workshop on the Future of Web Search, Barcelona, Spain, May 2006
- Google, Mountain View, USA, August 2006
- Yahoo! Research, Sunnyvale, USA, August 2006
- Microsoft Research, Redmond, USA, August 2006
- Tutorial at International Conference on Very Large Data Bases (VLDB), Seoul, Korea, September 2006
- Keynote, International Conference on Management of Data (COMAD), Delhi, India, December 2006
- Keynote, International Workshop on Ranking in Databases (DBRank), Istanbul, Turkey, April 2007
- Keynote, ACM SIGMOD Conference, Beijing, China, June 2007

16.8.5 Other Academic Activities

Gerhard Weikum:

- Member of the Board of Trustees and President of the VLDB Endowment
- Member of the ACM SIGMOD Advisory Board
- Member of the ACM SIGMOD Awards Committee
- Member of the Steering Committee of the Biennial Conference on Innovative Data Systems Research (CIDR)
- Program Committee Chair of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, USA, January 2007
- Member of the Best Paper Award Committee for the 23rd International Conference on Data Engineering (ICDE), Istanbul, Turkey, April 2007
- Member of the Best Paper Award Committee for the ACM SIGMOD International Conference on Management of Data, Chicago, USA, June 2006
- Member of the IEEE Data Engineering Working Group on Self-Managing Database Systems
- Member of the Scientific Advisory Board of the Swiss National Center of Competence in Research on "Mobile Information and Communication Systems" (MICS)
- Steering Committee Member of the German Computer Society Special Interest Area on Database and Information Systems (GI Fachbereich Datenbanken und Informationssystems (DBIS))
- Steering Committee Member of the German Computer Society Special Interest Group on Information Retrieval
- Spokesperson of the International Max Planck Research School for Computer Science (IMPRS-CS)
- Member of Search Committees at Saarland University and Hasso-Plattner Institute Potsdam
- Member of the Scientific Directorate of Schloss Dagstuhl (IBFI)
- Managing Director of the Max Planck Institute for Informatics
- Member of the Max Planck Society's Advisory Committee for IT Projects (BAR)
- Chair of the institute-internal committee for IT projects (HI-BAR)
- Member of the Max Planck Society's Steering Committee for Scientific Information (LA sInfo)

16.9 Teaching Activities

Summer Semester 2005

Courses:

Information Systems (R. Schenkel, G. Weikum)

Winter Semester 2005/2006

Courses:
Information Retrieval and Data Mining (G. Weikum)

Summer Semester 2006

Seminars:
Distributed Systems (T. Neumann)

Winter Semester 2006/2007

Courses:
Query Optimization (T. Neumann)

Diploma and Master's Theses

- Mohammad Alrifai: *Load Awareness in FliX*, October 2005
- Sebastian Blohm: *Exploiting Organizational Information in Enterprise Text Search*, November 2005
- Pleng Chirawatkul: *Structured Peer-to-Peer Search to build a Bibliographic Paper Recommendation System*, November 2006
- Munwar Hussain: *Language Modeling Based Passage Retrieval for Question Answering Systems*, September 2005
- Yasir Iqbal: *Flexible Backing-off Strategies for HMM based Named Entity Recognition*, April 2006
- Andreas Kaster: *Automatische Dokumentklassifikation mittels linguistischer und stilistischer Features*, April 2005
- Michael Koenig: *Effektives und effizientes Autoritätsranking von Webinhalten durch kontextspezifische Analyse und Optimierung der Konnektivität*, May 2005
- Andreas Martin: *Datenbereinigung und Datenintegration im Umweltsektor*, April 2005
- Gerard de Melo: *Multilingual Text Classification using Ontologies*, Johann Wolfgang Goethe-Universität Frankfurt am Main, December 2006
- Anna Moleda: *Probabilistic Scheduling for Top-k Query Processing*, August 2005
- Christian Nicolaus: *Echtzeitdatenerfassung und Datenkonsolidierung zur Qualitätssicherung einer Getriebefertigungsstrecke*, August 2005
- Odysseas Papapetrou: *On the Usage of Global Document Occurrences in Peer-to-Peer Information Systems*, October 2005
- Sebastian Parkitny: *A Comparative Study of Pub/Sub Methods in Structured P2P Networks*, October 2006
- Corinna Richter: *Klassifikation von Werkstoffgefügebildern aufgrund teilchenbasierter Kenngrößen*, March 2005

- Dennis Schade: *Effizientes Routing in Peer-to-Peer Netzwerken*, March 2005
- Thorsten Schneider: *Dynamics in Large-Scale Peer-to-Peer-Networks*, October 2006
- Isabell Schu: *Time-series Rule Discovery on Gene Expression Data*, March 2006
- Fabian Suchanek: *Ontological Reasoning for Natural Language Understanding*, March 2005

16.10 Dissertations, Habilitations, Offers, Awards

16.10.1 Dissertations

Completed:

Jens Graupmann: The SphereSearch Engine for Graph-Based Retrieval of Heterogeneous, Semistructured Data (in German), May 2006.

German Shegalov: Integrated Data, Message, and Process Recovery for Failure Masking in Web Services, July 2005.

Stefan Siersdorfer: Combination Methods for Automatic Document Organization, December 2005.

Sergej Sizov: Automatic Generation of Thematically Focused Information Portals from Web Data, December 2005.

Martin Theobald: Efficient Top-k Query Processing for Text, Semistructured, and Structured Data, May 2006.

In preparation:

Ralitsa Angelova: Neighborhood–Conscious Hypertext Categorization

Matthias Bender: Advanced Methods for Query Routing in Peer-to-Peer Information Retrieval

Klaus Berberich: Efficient Time-Travel Search over Web Archives

Andreas Broschart: Towards Top-k-Aware Graph Retrieval

Tom Crecelius: Social P2P Web Search

Gerard de Melo: Multilingual Text Classification Using Ontologies

Georgiana Ifrim: A Statistical Learning Approach to Concept-Based Document Classification

Gjergji Kasneci: Search with Entities and Relationships

Julia Luxenburger: Query-log Induced Implicit Feedback for Information Retrieval and Ranking

Sebastian Michel: Top-k Aggregation Queries in Large-Scale Distributed Systems

Hanglin Pan: Feedback-Driven Query Refinement in Ranked XML Retrieval

Josiane Xavier Parreira: Computing Global Authority Scores in a Peer-to-Peer Network

Fabian Suchanek: Information Extraction for Ontology Learning

Christian Zimmer: Approximate Information Filtering in Peer-to-Peer Networks

16.10.2 Awards

- *Ralitsa Angelova*, Best Poster Award at the 3rd British Computer Science Society Summer School on Pattern Recognition
- *Matthias Bender, Sebastian Michel, Gerhard Weikum, Christian Zimmer*, Best Interdisciplinary Paper Award at the 15th International Conference on Information and Knowledge Management (CIKM), Arlington, USA, 2006, for the paper *Discovering and exploiting keyword and attribute-value co-occurrences to improve P2P routing indices* (with Nikos Ntarmos and Peter Triantafillou)
- *Josiane Xavier Parreira, Sebastian Michel, Gerhard Weikum*, Selected for Best-of-VLDB'06 Special issue of the VLDB Journal, for the paper *Efficient and Decentralized PageRank Approximation in a Peer-to-Peer Web Search Network* (with Debora Donato)
- *Fabian Suchanek*, Günter-Hotz-Medal for outstanding Master Thesis, 2005
- *Martin Theobald*, Dissertation Award of the DBIS Section of the German Computer Society, 2007
- *Gerhard Weikum*, Fellow of the ACM, 2005

16.11 Grants and Cooperations

16.11.1 Projects Funded by the European Union

Integrated Project DELIS

DELIS (Dynamically Evolving Large-Scale Information Systems) is an Integrated Project under the 6th Framework Program of the European Union. It involves 20, mostly academic, partners and aims at foundations for the analysis and self-organization of complex systems such as peer-to-peer information sharing. DELIS is one of four EU projects funded under the Complex Systems Initiative. The institute participates with two of its groups, D1 and D5. The project has, to some extent, a catalytic function in stimulating joint research between the two groups.

D5 coordinates one of the six subprojects of DELIS, namely, SP6 on Data Management, Search, and Mining on Internet-Scale, Dynamically Evolving Peer-to-Peer Networks. It aims at developing foundations for collaborative Web search, with Google-style functionality but provided in a completely decentralized and self-organizing manner. Each peer has a full-fledged search engine and a local index, built by its own crawling and tailored to the user's personal interests. Peers collaborate by routing queries to thematically relevant peers and forming a dynamically evolving "semantic overlay network". In contrast to a centralized search engine running on a high-end server farm and geared for high throughput of simple mass-user queries, DELIS pursues advanced information retrieval models and algorithms like concept-based search of different flavors. D1 and D5 have been collaborating on this theme, in particular on the efficiency challenges for such advanced search capabilities. Other key issues in DELIS SP6 include distributed indexing, efficient overlay networks, information dissemination, incentive mechanisms, and a deeper understanding of benefit/cost tradeoffs and optimizations.

Network of Excellence DELOS

DELOS is a network of excellence on digital library systems. It currently involves about 50 partners, and is organized into 8 thematic clusters. D5 participates in clusters WP1 and WP2 on digital library system architecture and personalized access.

The main issues pursued by D5 in this context are query routing and query processing over federations of digital libraries and the exploitation of personal profiles for routing strategies. D5 collaborates 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 has been started on analyzing access and behavior information of digital-library users. This work is centered around the action logs of the TEL portal to national libraries in Europe. It involves collaboration with TEL (The European Library), the University of Padova, and the University of Athens.

SAPIR

SAPIR (Search on Audio-Visual Content using Peer-to-Peer Information Retrieval) is a STREP project (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, MPI-INF). 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. MPI-INF 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.

16.11.2 Projects Funded by the BMBF

P2E2

The P2E2 project has been a joint project of several industry partners and two academic partners on peer-to-peer-style workflow management, funded by the German Ministery of Science and Education and running from October 2003 to December 2006. The industrial partners were Carnot AG, abaXX Technology AG, otris Software AG, IDS Scheer AG; the second academic partner was the Institute for Information Systems at the German Research Center for Artificial Intelligence (DFKI). The project has developed mechanisms and strategies orchestrating workflows that span enterprise boundaries. Such processes should be run in a decentralized manner but, at the same time, it should be possible to monitor their

progress from a conceptually global viewpoint. This included studying peer-to-peer methods for data gathering and distributed queries.

16.11.3 Projects Funded by the DFG

CLASSIX

The CLASSIX project has been a joint project of D5 and the information retrieval group of the University of Duisburg headed by Prof. Norbert Fuhr. The project started in February 2002 and ended in Feburary 2007. CLASSIX has developed new methods for intelligently searching and organizing XML data. It has thus addressed a strategically important slice of the theme of integrating database-system (DB) and information-retrieval (IR) technologies, by providing concepts, models, algorithms, and system know-how about XML IR. effective, techniques for information retrieval and automatic classification of XML data. Based on the groups' extensive systems work (i.e., the prototype systems HyREX and Daffodil on the Duisburg side and BINGO!, XXL, and TopX on the MPI-INF side) the project has investigated query and ranking models, query evaluation and indexing techniques, the integration of ontological metadata, relevance feedback methods, and automatic classification methods, with experimental evaluation based on the INEX and other benchmarks. Both partners have played important roles in the INEX benchmarking initiative: the Duisburg group has been one of the organizers, the MPI-INF group has provided the reference engine for topic development.

A particularly noteworthy joint result of the two groups is the coupling of the BINGO! focused crawler and the digital library mediator Daffodil, leading to added value to a federated digital library environment where advanced users can manage personal folders with annotations, access multiple libraries in a seamless manner, and enhance library data by Web data compiled via focused crawling.

Graduiertenkolleg "Quality Guarantees for Computer Systems"

The group actively participates in the Saarland University's graduate studies program (Graduiertenkolleg) on "Quality Guarantees for Computer Systems". Currently, two doctoral students receive scholarships from this program.

16.11.4 Cooperations with Industry

In addition to the industrial partners in the EU and BMBF projects, D5 maintains loose collaborations (e.g., through Diploma or Master theses) with various local and non-local companies, including SAP, IBM, Microsoft, Yahoo!, and the local startup companies Deep Web and Consistec.

16.12 Publications

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, May 2005.

[2] S. Amer-Yahia, P. Case, T. Roelleke, J. Shanmugasundaram, and G. Weikum. Report on the DB/IR panel at SIGMOD 2005. *SIGMOD Record*, 34(4):71–74, December 2005.

[3] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. "To infinity and beyond": P2P web search with Minerva and Minerva∞. In R. Baldoni, G. Cortese, F. Davide, and A. Melpignano, eds., *Global Data Management*, vol. 8, ch. Applications, pp. 301–323. IOSPress, Amsterdam, The Netherlands, 2006.

[4] M. Bender, S. Michel, G. Weikum, and C. Zimmer. Das MINERVA-Projekt: Datenbankselektion für Peer-to-Peer-Websuche. *Informatik - Forschung und Entwicklung*, 20(3):152 – 166, December 2005.

[5] K. Berberich, M. Vazirgiannis, and G. Weikum. Time-aware authority ranking. *Internet Mathematics*, 2(3):301–332, 2006.

[6] S. Chaudhuri, G. Das, V. Hristidis, and G. Weikum. Probabilistic information retrieval approach for ranking of database query results. *ACM Transactions on Database Systems*, 31(3):1134–1168, September 2006.

[7] Y. Ioannidis, D. Maier, S. Abiteboul, P. Buneman, S. Davidson, E. Fox, A. Halevy, C. Knoblock, F. Rabitti, H. Schek, and G. Weikum. Digital library information-technology infrastructures. *International Journal on Digital Libraries*, 5(4):266–274, August 2005. Special section on NSF/EU-DELOS working groups.

[8] C. Jones, D. Lomet, A. Romanovsky, G. Weikum, A. Fekete, M.-C. Gaudel, H. F. Korth, R. de Lemos, J. E. B. Moss, R. Rajwar, K. Ramamritham, B. Randell, and L. Rodrigues. The atomic manifesto: a story in four quarks. *SIGMOD Record*, 34(1):63–69, March 2005.

[9] C. Jones, D. Lomet, A. Romanovsky, G. Weikum, A. Fekete, M.-C. Gaudel, H. F. Korth, R. de Lemos, J. E. B. Moss, R. Rajwar, K. Ramamritham, B. Randell, and L. Rodrigues. The atomic manifesto: a story in four quarks. *SIGOPS Operating Systems Review*, 39(2):41–46, April 2005.

[10] C. B. Jones, D. B. Lomet, A. B. Romanovsky, and G. Weikum. The atomic manifesto. *Journal of Universal Computer Science*, 11(5):636–651, May 2005.

[11] J. X. Parreira, S. Michel, and G. Weikum. p2pDating: Real life inspired semantic overlay networks for web search. *Information Processing & Management*, 43(3):643–664, 2007.

[12] K. Roberts, F. Mücklich, and G. Weikum. Untersuchungen zur automatischen Klassifikation von Lamellengraphit mit Hilfe des Stützvektorverfahrens (Examinations on the Automatic Classification of Lamellar Graphite Using the Support Vector Machine). *Praktische Metallographie = Practical metallography : international journal on materialographic preparation, imaging and analysis of microstructures*, 42(8):396–410, August 2005.

[13] R. Schenkel, A. Theobald, and G. Weikum. Semantic similarity search on semistructured data with the XXL search engine. *Information Retrieval*, 8(4):521–545, December 2005.

[14] G. Weikum. The database research group at the Max-Planck Institute for Informatics. *SIGMOD Record*, 35(3):42–47, September 2006.

Conference articles

[1] R. Angelova and S. Siersdorfer. A neighborhood-based approach for clustering of linked document collections. In P. S. Yu, V. J. Tsotras, E. A. Fox, and B. Liu, eds., *Proceedings of the Conference on Information and Knowledge Management, CIKM 2006*, Arlington, VA, USA, 2006, pp. 778–779. ACM.

[2] R. Angelova and G. Weikum. Graph-based text classification: Learn from your neighbors. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Jaervelin, eds., *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, 2006, pp. 485–492. ACM. Acceptance ratio 1:5.

[3] 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.

[4] 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.

[5] P. Baumgartner and F. M. Suchanek. Automated reasoning support for first order ontologies. In J. J. Alferes, J. Bailey, W. May, and U. Schwertel, eds., *Principles and Practice of Semantic Web Reasoning, 4th International Workshop, PPSWR 2006*, Budva, Montenegro, 2006, *LNCS 4187*, pp. 18–32. Springer. Acceptance ratio 1:5.

[6] S. Bedathur and J. Haritsa. Search-optimized suffix-tree storage for biological applications. In D. A. Bader, M. Parashar, S. Varadarajan, and V. K. Prasanna, eds., *12th IEEE International Conference on High Performance Computing (HiPC)*, Goa, India, October 2006, *LNCS 3769*, pp. 29–39. Springer.

[7] M. Bender, N. Fuhr, Y. Ioannidis, D. Kossman, H.-J. Schek, G. Weikum, P. Zezula, and C. Zimmer. Personalized query routing in peer-to-peer federations of digital libraries. In C. Thanos, ed., *DELOS Research Activities 2006*, pp. 29–32. IST-CNT, Pisa, Italy, 2007.

[8] M. Bender, Y. Ioannidis, H. Nottelmann, H.-J. Schek, G. Weikum, P. Zezula, and C. Zimmer. Personalized query routing in peer-to-peer federations of digital libraries. In C. Thanos, ed., *DELOS Research Activities 2005*, pp. 17–18. ISTI-CNR, Pisa, Italy, 2007.

[9] M. Bender, S. Michel, S. Parkitny, and G. Weikum. A comparative study of pub/sub methods in structured P2P networks (to be published). In S. Bergamaschi, S. Joseph, J.-H. Morin, and G. Moro, eds., *Fourth International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2006)*, Seoul, South Korea, 2006. Also published at Delos.

[10] M. Bender, S. Michel, P. Triantafillou, and G. Weikum. Global document frequency estimation in peer-to-peer web search. In D. Zhou, ed., *9th International Workshop on the Web and Databases (WebDB 2006) @ SIGMOD2006*, Chicago, USA, 2006, pp. 69–74. Acceptance Ratio 1:4.

[11] 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 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brazil, 2005, pp. 67–74. ACM. Acceptance ratio 1:5.

[12] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. P2P content search: Give the web back to the people. In *5th International Workshop on Peer-to-Peer Systems (IPTPS 2006)*, Santa Barbara, US, 2006. Acceptance ratio: 1:4.

[13] M. Bender, S. Michel, and G. Weikum. P2P directories for distributed web search: From each according to his ability, to each according to his needs. In R. S. Barga and X. Zhou, eds., *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*, Atlanta, GA, USA, 2006, pp. 1–10. IEEE.

[14] M. Bender, S. Michel, G. Weikum, and C. Zimmer. Challenges of distributed search across digital libraries. In G. Weikum, Y. Ioannidis, and H.-J. Schek, eds., *Proceedings of the 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems (System Architecture & Information Access)*, Dagstuhl, Germany, 2005, pp. 55–59. Information Society Technologies.

[15] M. Bender, S. Michel, G. Weikum, and C. Zimmer. The MINERVA project: Database selection in the context of P2P search. In G. Vossen, F. Leymann, P. C. Lockemann, and W. Stucky, eds., *Datenbanksysteme in Business, Technologie und Web (BTW), 11. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS)*, Karlsruhe, Germany, March 2005, *Lecture Notes in Informatics*, vol. 65, pp. 125–144. Gesellschaft für Informatik. Acceptance ratio 1:3.

[16] M. Bender, S. Michel, C. Zimmer, and G. Weikum. Towards collaborative search in digital libraries using peer-to-peer technology. In C. Türker, M. Agosti, and H.-J. Schek, eds., *Peer-to-peer, grid, and service-orientation in digital library architectures, 6th Thematic Workshop of the EU Network of Excellence DELOS*, Cagliari, Italy, August 2005, *LNCS 3664*, pp. 80–95. Springer. Selected, Revised Papers.

[17] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, Netherlands, July 23-27, 2007*, Amsterdam, Netherlands, 2007. ACM.

[18] K. Berberich, S. Bedathur, M. Vazirgiannis, and G. Weikum. Comparing apples and oranges: Normalized pagerank for evolving graphs. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, May 2007. ACM.

[19] 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.

[20] K. Berberich, S. J. Bedathur, M. Vazirgiannis, and G. Weikum. Buzzrank ... and the trend is your friend. In *Proceedings of the 15th International World Wide Web Conference (WWW 2006)*, Edinburgh, UK, 2006, pp. 937–938. ACM.

[21] K. Berberich, S. J. Bedathur, and G. Weikum. Rank synopses for efficient time travel on the web graph. 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. 864–865. ACM.

[22] K. Berberich, M. Koubarakis, C. Tryfonopoulos, G. Weikum, and C. Zimmer. MAPS: Approximate publish/subscribe functionality in peer-to-peer networks. In *ADPUC '06: Proceedings of the 1st International Workshop on Advanced Data Processing in Ubiquitous Computing (ADPUC 2006)*, Melbourne, Australia, 2006, *ACM International Conference Proceeding Series*, vol. 181, pp. 1–6. ACM.

[23] A. Broschart. Efficient integration of proximity for text, semistructured and graph retrieval. In *SIGIR 2007 Doctoral Consortium*, Amsterdam, The Netherlands, 2007.

[24] S. Chaudhuri, R. Ramakrishnan, and G. Weikum. Integrating DB and IR technologies: What is the sound of one hand clapping? In M. Stonebraker, G. Weikum, and D. DeWitt, eds., *Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR 05)*, Asilomar, CA, USA, 2005, pp. 1–12. VLDB. Acceptance ratio 1:3.

[25] S. Chaudhuri and G. Weikum. Foundations of automated database tuning (tutorial). In J. Widom, F. Özcan, and R. Chrikova, eds., *SIGMOD 2005, proceedings of the ACM SIGMOD International Conference on Management of Date*, Baltimore, USA, 2005, pp. 964–965. ACM.

[26] S. Chaudhuri and G. Weikum. Foundations of automated database tuning (tutorial). 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. 1265–1265. ACM.

[27] S. Chaudhuri and G. Weikum. Foundations of automated database tuning (tutorial). In L. Liu, A. Reuter, K.-Y. Whang, and J. Zhang, eds., *Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006)*, Atlanta, GA, USA, 2006, pp. 1–1. IEEE.

[28] S. Chernov, P. Serdyukov, M. Bender, S. Michel, G. Weikum, and C. Zimmer. Database selection and result merging in P2P web search. In *3rd International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2005)*, Trondheim, Norway, 2005, pp. 1–14. Acceptance rate 1:3.

[29] J. Graupmann, J. Cai, and R. Schenkel. Automatic query refinement using mined semantic relations. In *International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)*, Tokyo, Japan, 2005, pp. 205–213. IEEE. Acceptance ratio 1:4.

[30] J. Graupmann and R. Schenkel. The light-weight semantic web: Integrating information extraction and information retrieval for heterogeneous environments. In *SIGIR 2005 Workshop on Heterogeneous and Distributed Information Retrieval (HDIR)*, Salvador, Brazil, August 2005, pp. 1–8. ACM.

[31] J. Graupmann and R. Schenkel. GeoSphereSearch: Context-aware geographic web search. In R. Purves and C. Jones, eds., *3rd Workshop on Geographic Information Retrieval*, Seattle, WA, USA, 2006, pp. 1–4.

[32] 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.

[33] 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.

[34] 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.

[35] N. Kammenhuber, J. Luxenburger, A. Feldmann, and G. Weikum. Web search clickstreams. In J. M. Almeida, V. A. F. Almeida, and P. Barford, eds., *Proceedings of the 6th ACM SIGCOMM on Internet measurement (IMC '06)*, Rio de Janeiro, Brazil, 2006, pp. 245–250. ACM.

[36] G. Kasneci and T. Schwentick. The complexity of reasoning about pattern-based XML schemas. In *26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2007)*, Beijing, China, 2007. ACM press.

[37] G. Kasneci, F. Suchanek, M. Ramanath, and G. Weikum. How NAGA uncoils: Searching with entities and relations. In *16th International World Wide Web Conference (WWW 2007)*, New York, NY, USA, 2007. ACM Press.

[38] A. Kaster, S. Siersdorfer, and G. Weikum. Combining text and linguistic document representations for authorship attribution. In S. Argamon, J. Karlgren, and J. G. Shanahan, eds., *Workshop Stylistic Analysis of Text for Information Access, 28th International SIGIR*, Salvador, Brazil, August 2005, vol. 1, pp. 27–35. ACM.

[39] A. Linari and G. Weikum. Efficient peer-to-peer semantic overlay networks based on statistical language models. In *P2PIR '06: Proceedings of the International Workshop on Information Retrieval in Peer-to-peer Networks*, Arlington, Virginia, USA, 2006, pp. 9–16. ACM.

[40] D. B. Lomet, R. S. Barga, M. F. Mokbel, G. Shegalov, R. Wang, and Y. Zhu. Transaction time support inside a database engine. In L. Liu, A. Reuter, K.-Y. Whang, and J. Zhang, eds., *Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006)*, Atlanta, USA, 2006, pp. 1–12. IEEE.

[41] J. Luxenburger and G. Weikum. Exploiting community behavior for enhanced link analysis and web search. In D. Zhou, ed., *Proceedings of the 9th International Workshop on the Web and Databases (WebDB 2006)*, Chicago, Illinois, USA, 2006, pp. 14–19.

[42] D. Mavroeidis, G. Tsatsaronis, M. Vazirgiannis, M. Theobald, and G. Weikum. Word sense disambiguation for exploiting hierarchical thesauri in text classification. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and G. Joao, eds., *Knowledge discovery in databases: PKDD 2005, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Porto, Portugal, 2005, *LNCS 3721*, pp. 181–192. Springer.

[43] G. de Melo and S. Siersdorfer. Multilingual text classification using ontologies. In G. Amati, ed., *Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007)*, Rome, Italy, 2007, *LNCS 4425*. Springer. Acceptance Ratio 1:4.

[44] 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.

[45] S. Michel, M. Bender, P. Triantafillou, and G. Weikum. IQN routing: Integrating quality and novelty in P2P querying and ranking. In Y. Ioannidis, M. H. Scholl, J. W. Schmidt, F. Matthes, M. Hatzopoulos, K. Boehm, A. Kemper, T. Grust, and C. Boehm, eds., *Advances in Database Technology - EDBT 2006: 10th International Conference on Extending Database Technology*, Munich, Germany, March 2006, *LNCS 3896*, pp. 149–166. Springer.

[46] S. Michel, P. Triantafillou, and G. Weikum. KLEE: A framework for distributed top-k query algorithms. 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. 637–648. ACM. Acceptance ratio 1:6.

[47] S. Michel, P. Triantafillou, and G. Weikum. Minerva∞: A scalable efficient peer-to-peer search engine. In G. Alonso, ed., *Middleware 2005, ACM, IFIP, USENIX 6th International Middleware Conference*, Grenoble, France, 2005, *LNCS 3790*, pp. 60–81. Springer.

[48] 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, September 2006, pp. 930–941. ACM. Acceptance ratio 1:7.

[49] T. Neumann, M. Bender, S. Michel, and G. Weikum. A reproducible benchmark for P2P retrieval. In P. Bonnet and I. Manolescu, eds., *Proceedings of the 1st International Workshop on Performance and Evaluation of Data Management Systems, ExpDB 2006, in cooperation with ACM SIGMOD*, Chicago, Illinois, USA, 2006, pp. 1–8. ACM.

[50] T. Neumann, S. Helmer, and G. Moerkotte. On the optimal ordering of maps, selections, and joins under factorization. In D. J. Bell and J. Hong, eds., *Flexible and efficient information handling, 23rd British National Conference on Databases, BNCOD 23*, Belfast,Northern Ireland, UK, September 2006, *LNCS 4042*, pp. 115–126. Springer.

[51] 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.

[52] N. Ntarmos, P. Triantafillou, and G. Weikum. Counting at large: Efficient cardinality estimation in internet-scale data networks. In L. Liu, A. Reuter, K.-Y. Whang, and J. Zhang, eds., *Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006)*, Atlanta, GA, USA, 2006, pp. 1–10. IEEE.

[53] O. Papapetrou, S. Michel, M. Bender, and G. Weikum. On the usage of global document occurrences in peer-to-peer information systems. In R. Meersman, Z. Tari, M.-S. Hacid, J. Mylopoulos, B. Pernici, Ö. Babaoglu, H.-A. Jacobsen, J. P. Loyall, M. Kifer, and S. Spaccapietra, eds., *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005*, Agia Napa, Cyprus, 2005, *LNCS 3760*, pp. 310–328. Springer.

[54] 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. .

[55] 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.

[56] J. X. Parreira, S. Michel, and M. Bender. Size doesn't always matter: Exploiting pagerank for query routing in distributed IR. In *P2PIR '06: Proceedings of the International Workshop on Information Retrieval in Peer-to-peer Networks*, Arlington, USA, 2006, pp. 25–32. ACM.

[57] J. X. Parreira, S. Michel, and G. Weikum. p2pDating: Real life inspired semantic overlay networks for web search. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, eds., *Proceedings of the Workshop on Heterogeneous and Distributed Information Retrieval*, Salvador Bahia, Brazil, 2005, pp. 1–8. ACM.

[58] J. X. Parreira and G. Weikum. JXP: Global authority scores in a P2P network. In A. Doan, F. Neven, R. McCann, and G. Jan Bex, eds., *Proceedings of the 8th International Workshop on the Web & Databases (WebDB 2005) collocated with ACM SIGMOD/PODS 2005*, Baltimore, Maryland, USA, 2005, pp. 31–36. ACM.

[59] R. Schenkel, F. M. 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.

[60] R. Schenkel, A. Theobald, and G. Weikum. Efficient creation and incremental maintenance of the HOPI index for complex XML document collections. In *Proceedings of the 21st International Conference on Data Engineering, ICDE 2005*, Tokyo, Japan, 2005, pp. 360–371. IEEE. Acceptance ratio 1:8.

[61] R. Schenkel and M. Theobald. Feedback-driven structural query expansion for ranked retrieval of XML data. In Y. Ioannidis, M. H. Scholl, J. W. Schmidt, F. Matthes, M. Hatzopoulos, K. Boehm, A. Kemper, T. Grust, and C. Boehm, eds., *Advances in Database Technology - EDBT 2006: 10th International Conference on Extending Database Technology*, Munich, Germany, 2006, *LNCS 3896*, pp. 331–348. Springer. Acceptance ratio 1:6.

[62] R. Schenkel and M. Theobald. Relevance feedback for structural query expansion. In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, eds., *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, Dagstuhl Castle, Germany, 2006, *LNCS 3977*, pp. 344–357. Springer.

[63] 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.

[64] G. Shegalov, G. Weikum, and K. Berberich. Unstoppable stateful PHP web services. In K. Aberer, Z. Peng, E. A. Rundensteiner, Y. Zhang, and X. Li, eds., *Web Information Systems - WISE 2006, 7th International Conference on Web Information Systems Engineering*, Wuhan, China, 2006, *LNCS 4255*, pp. 132–143. Springer.

[65] S. Siersdorfer and S. Sizov. Automatic document organization in a P2P environment. 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. 265–276. Springer.

[66] S. Siersdorfer and G. Weikum. Automated retraining methods for document classification and their parameter tuning. In A. H. H. Ngu, M. Kitsuregawa, E. J. Neuhold, J.-Y. Chung, and Q. Z. Sheng, eds., *Web information systems engineering - WISE 2005, 6th International Conference on Web Information Systems Engineering*, New York, USA, 2005, *LNCS 3806*, pp. 478–486. Springer.

[67] S. Siersdorfer and G. Weikum. Using restrictive classification and meta classification for junk elimination. In D. Losada and J. M. Fernandez-Luna, eds., *Advances in information retrieval, 27th European Conference on IR Research, ECIR 2005*, Santiago de Compostela, Spain, March 2005, *LNCS 3408*, pp. 287–299. Springer. Acceptance ratio 1:4.

[68] F. M. 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.

[69] F. M. Suchanek, G. Ifrim, and G. Weikum. LEILA: Learning to extract information by linguistic analysis. In P. Buitelaar, P. Cimiano, and B. Loos, eds., *Proceedings of the 2nd Workshop on Ontology Learning and Population (OLP2) @COLING/ACL 2006*, Sydney, Australia, 2006, pp. 18–25. ACL. Acceptance Rate: 8/19=42

[70] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge - unifying WordNet and Wikipedia. In *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007. IW3C2.

[71] 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., *Preproceedings of the 5th International Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2006)*, Schloß Dagstuhl, Germany, 2006, pp. 140–149.

[72] 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.

[73] 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.

[74] M. Theobald, R. Schenkel, and G. Weikum. TopX & XXL at INEX 2005 (ad-hoc track). In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, eds., *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005*, Dagstuhl Castle, Germany, 2006, *LNCS 3977*, pp. 282–295. Springer.

[75] 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.

[76] G. Weikum. Informations- und Wissensmanagement im Jahr 2025: BTW allez oder BTW passée. In G. Vossen, F. Leymann, P. C. Lockemann, and W. Stucky, eds., *Datenbanksysteme in Business, Technologie und Web (BTW), 11. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS)*, Karlsruhe, Germany, March 2005, *Lecture Notes in Informatics*, vol. 65, pp. 29–29. Gesellschaft für Informatik.

[77] G. Weikum, H. Bast, G. Canright, D. Hales, C. Schindelhauer, and P. Triantafillou. Towards self-organizing query routing and processing for peer-to-peer web search. In V. Damerov, ed., *Workshop on Peer-to-peer Data Management in the Complex Systems Perspective*, Paris, France, November 2005, pp. 7–24. University of Paderborn, Heinz Nixdorf Institute.

Demonstrations (in Referreed and Selective Demo Programs at Conferences)

[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 07)*, Asilomar, USA, January 2007, pp. 164–168. www.crdrdb.org.

[2] 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.

[3] 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. IEEE Computer Society.

[4] D. Lomet, R. Barga, M. Mokbel, G. Shegalov, R. Wang, and Y. Zhu. Immortal DB: transaction time support for SQL server (demo). In J. Widom, F. Özcan, and R. Chrikova, eds., *SIGMOD 2005, Proceedings of the ACM SIGMOD International Conference on Management of Data*, Baltimore, Maryland, USA, 2005, pp. 939–941. ACM.

[5] S. Michel, M. Bender, P. Triantafillou, G. Weikum, and C. Zimmer. P2P web search with MINERVA: How do you want to search tomorrow? (demo), 2005.

[6] G. Shegalov and G. Weikum. EOS^2: Unstoppable stateful PHP (demo). 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. 1223–1226. ACM.

[7] 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. ACM.

Theses

[1] G. Aalam. Georeferenzierung von Suchergebnissen auf Basis von annotierten Daten und Geoinformationssystemen. Bachelor thesis, Universität des Saarlandes, March 2007.

[2] M. Al-Rifai. Load awareness in flix. Masters thesis, Universität des Saarlandes, October 2005.

[3] S. Blohm. Exploiting organizational information in enterprise text search. Masters thesis, Universität des Saarlandes, November 2005.

[4] S. Chernov. Result merging in a peer-to-peer Web search engine. Masters thesis, Universität des Saarlandes, March 2005.

[5] P. Chirawatkul. Structured peer-to-peer search to build a bibliographic paper recommendation system. Masters thesis, Universität des Saarlandes, November 2006.

[6] D. K. Fuchs. Verteilte Annotation von Dokumenten - Entwurf und Implementierung eines modularen Annotationsservers auf Basis von NLP-Technologien. Bachelor thesis, Universität des Saarlandes, June 2006.

[7] F. S. Fuchs. Verteilte Annotation von Dokumenten - Architekturentwurf und Implementierung eines Frameworks auf J2EE-Basis. Bachelor thesis, Universität des Saarlandes, June 2006.

[8] J. Graupmann. *Die Sphere-Search-Suchmaschine zur Graphbasierten Suche auf Heterogenen, Semistrukturierten Daten*. Phd thesis, Universität des Saarlandes, May 2006.

[9] F. Guo. Adaptation of the focused web crawler bingo! for the Minerva project. Bachelor thesis, Universität des Saarlandes, January 2006.

[10] M. Hussain. Language modeling based passage retrieval for question answering systems. Masters thesis, Universität des Saarlandes, September 2005.

[11] G. Ifrim. A bayesian learning approach to concept-based document classification. Masters thesis, Universität des Saarlandes, March 2005.

[12] Y. Iqbal. Flexible backing-off strategies for hmm based named entity recognition. Masters thesis, Universität des Saarlandes, April 2006.

[13] A. Kaster. Automatische Dokumentklassifikation mittels linguistischer und stilistischer Features. Masters thesis, Universität des Saarlandes, April 2005.

[14] M. Koenig. Effektives und effizientes Autoritätsranking von Webinhalten durch kontextspezifische Analyse und Optimierung der Konnektivität. Masters thesis, Universität des Saarlandes, May 2005.

[15] N. Kozlova. Automatic ontology extraction for document classification. Masters thesis, Universität des Saarlandes, March 2005.

[16] T. Mangel. Query-driven Term Correlations for Advanced P2P Query Routing. Bachelor thesis, Universität des Saarlandes, April 2006.

[17] A. Martin. Datenbereinigung und Datenintegration im Umweltsektor. Masters thesis, Universität des Saarlandes, April 2005.

[18] A. Moleda. Probabilistic scheduling for top-k query processing. Masters thesis, Universität des Saarlandes, August 2005.

[19] C. Nicolaus. Echtzeitdatenerfassung und Datenkonsolidierung zur Qualitätssicherung einer Getriebefertigungsstrecke. Masters thesis, Universität des Saarlandes, August 2005.

[20] M. Nicolay. Design und Implementierung eines Sync-ML-Clients für das c'man-Framework für Symbian-Geräte. Bachelor thesis, Universität des Saarlandes, March 2007.

[21] O. Papapetrou. On the usage of global document occurrences in peer-to-peer information systems. Masters thesis, Universität des Saarlandes, October 2005.

[22] S. Parkitny. A comparative study of pub/sub methods in structured p2p networks. Masters thesis, Universität des Saarlandes, October 2006. This work also served as a basis for the DBISP2P 2006 paper.

[23] A. Prohaska. Visualisierung einer verteilten Hashtabelle (CHORD). Bachelor thesis, Universität des Saarlandes, October 2005.

[24] C. Richter. Klassifikation von Werkstoffgefügebildern aufgrund teilchenbasierter Kenngrößen. Masters thesis, Universität des Saarlandes, March 2005.

[25] D. Schade. Effizientes Routing in Peer-to-Peer Netzwerken. Masters thesis, Universität des Saarlandes, March 2005.

[26] T. Schneider. Dynamics in large-scale peer-to-peer-networks. Masters thesis, Universität des Saarlandes, October 2006.

[27] M. Schreiber. Neighbourhood-conscious Record Linkage. Bachelor thesis, Universität des Saarlandes, June 2006.

[28] I. Schu. Time-series rule discovery on gene expression data. Masters thesis, Universität des Saarlandes, March 2006.

[29] P. Serdyukov. Query routing in peer-to-peer Web search. Masters thesis, Universität des Saarlandes, March 2005.

[30] G. Shegalov. *Integrated Data, Message, and Process Recovery for Failure Masking in Web Services.* Phd thesis, Universität des Saarlandes, July 2005.

[31] S. Siersdorfer. *Combination Methods for Automatic Document Organization.* Phd thesis, Universität des Saarlandes, December 2005.

[32] S. Sizov. *Automatic Generation of Thematically Focused Information Portals from Web Data.* Phd thesis, Universität des Saarlandes, December 2005.

[33] F. M. Suchanek. Ontological reasoning for natural language understanding. Masters thesis, Universität des Saarlandes, March 2005.

[34] M. Theobald. *Efficient Top-k Query Processing for Text, Semistructured, and Structured Data.* Phd thesis, Universität des Saarlandes, May 2006.

Technical reports

[1] R. Angelova and S. Siersdorfer. A neighborhood-based approach for clustering of linked document collections. Research Report MPI-I-2006-5-005, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, September 2006.

[2] H. Bast, D. Majumdar, R. Schenkel, M. Theobald, and G. Weikum. Io-top-k: Index-access optimized top-k query processing. Research Report MPI-I-2006-5-002, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, March 2006.

[3] M. Bender, S. Michel, G. Weikum, and P. Triantafilou. Overlap-aware global df estimation in distributed information retrieval systems. Research Report MPI-I-2006-5-001, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, January 2006.

[4] G. Ifrim, G. Kasneci, M. Ramanath, F. M. Suchanek, and G. Weikum. Naga: Searching and ranking knowledge. Research Report MPI-I-2007-5-001, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, March 2007.

[5] G. Kasnec, F. M. Suchanek, and G. Weikum. Yago - a core of semantic knowledge. Research Report MPI-I-2006-5-006, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, November 2006.

[6] S. Siersdorfer and G. Weikum. Automated retraining methods for document classification and their parameter tuning. Research Report MPI-I-2005-5-002, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, September 2005.

[7] F. Suchanek, G. Ifrim, and G. Weikum. Combining linguistic and statistical analysis to extract relations from web documents. Research Report MPI-I-2006-5-004, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, March 2006.

17 The Automation of Logic Group (RG1)

17.1 Personnel

Head of the Group

Priv. Doz. Dr. Christoph Weidenbach

Researchers

Thomas Hillenbrand[1]
Manuel Lamotte (February 2007–)
Priv. Doz. Dr. Viorica Sofronie-Stokkermans[1]
Dr. Uwe Waldmann[1]

PhD Students

Matthias Horbach (Juli 2006–)
Carsten Ihlemann[1]
Swen Jacobs[1]

Secretary

Roxane Wetzel

17.2 Visitors

In the time period from September 2005 to March 2007, the following researchers visited our group:

Anja Feldmann	27.03.06	Deutsche Telekom Laboratory, TU Berlin
David Basin	29.03.06	ETH Zürich
Peter Baumgartner	31.07.06	Universität Koblenz
Renate A. Schmidt	12.10.06	University of Manchester
Johannes Faber	12.12.06–14.12.06	C. v. O. Universität Oldenburg
Stefan Schulz	15.12.06	TU München

[1] Thomas Hillenbrand, Carsten Ihlemann, Swen Jacobs, Viorica Sofronie-Stokkermans, and Uwe Waldmann moved from Department 2 to Research Group 1 in September 2005. With the exception of Subsection 13.3.2, their work is reported in this chapter.

17.3 Group Organization

We are a small group that does not require special organizational mechanisms. A weekly internal group meeting is established for student talks, the preparation of conference talks, discussion and exchange of ideas, and to manage all operational issues.

All our PhD students have a dedicated mentor who, in addition to scientific support, takes care for the writing of a half yearly progress report and accompanies the student along the overall way towards his PhD. Currently, Viorica Sofronie-Stokkermans supports Carsten Ihleman and Swen Jacobs and Christoph Weidenbach supports Thomas Hillenbrand, Matthias Horbach and Manuel Lamotte.

17.4 Hierarchic and Modular Reasoning

Complex systems arise in a natural way in many areas such as mathematics, logic, computation, databases, or artificial intelligence. Usually one would like to perform the verification tasks for such systems in a modular way, by using existing tools as black-boxes for verifying the components. Such modular approaches are important both for theorem proving in combinations of logical theories (it can help in drastically reducing the number of steps in a proof), and for the verification of systems consisting of several components which interact (it can help in controlling the explosion of the state space). Hierarchic situations arise in two ways. Firstly, when a theory is extended by additional structure. Secondly, they can also arise inside one theory, when reasoning in a specific theory can be reduced to reasoning in a less complex subtheory. All aspects are reflected in our work. We analyzed possibilities of hierarchic and modular reasoning in combinations of theories and their applications to the deductive verification (e.g. of real time systems), as well as possibilities of modular verification of complex systems, and have started to investigate the reduction of inductive theorem proving to first-order saturation.

17.4.1 Theorem proving in complex theories

Investigators: Harald Ganzinger, Carsten Ihlemann, Swen Jacobs, Viorica Sofronie-Stokkermans, and Uwe Waldmann

Many problems in mathematics and computer science can be reduced to proving satisfiability of conjunctions of (ground) literals modulo a background theory. This theory can be a standard theory, the extension of a base theory with additional functions, or a combination of theories. It is therefore very important to find efficient methods for reasoning in complex theories. We developed methods for hierarchical reasoning in theory extensions and for modular reasoning in combinations of theories.

Hierarchical and modular superposition-based calculi

We first analyzed the possibility of modular reasoning in combinations of theories within a resolution framework [3, 4]. Our aim was to reduce the number of inference steps in proofs, e.g. by excluding inferences between clauses with symbols from different theories. We gave a modular superposition calculus for the combination of first-order theories involving both

total and partial functions. The modularity of the calculus is a consequence of the fact that all the inferences are pure – only involving clauses over the alphabet of either one, but not both, of the theories – when refuting goals represented by sets of pure formulae. The calculus is sound; it is shown to be also complete provided that functions that are not in the intersection of the component signatures are declared as partial. This result also means that if the unsatisfiability of a goal modulo the combined theory does not depend on the totality of the functions in the extensions, the inconsistency will be effectively found. Moreover, we considered a constraint superposition calculus for the case of hierarchical theories and showed that it has a related modularity property. Finally we identified cases where the partial models can always be made total so that modular superposition is also complete with respect to the standard (total function) semantics of the theories.

Hierarchical reasoning in theory extensions

The line of research started in [3, 4] was continued in subsequent work. In [11] we identified a class of theory extensions, which we called *local theory extensions*, for which a direct hierarchical reasoning method (in which the problems are reduced to reasoning in the base theory) is possible – without the need for using a superposition or resolution framework. The completeness of the method we used only relies on the locality of the extension. We showed that, for local theory extensions, checking the validity of a universal formula can be reduced to checking the validity of a set of clauses with respect to the base theory – for this, any specialized theorem prover for the base theory can be used as a "black box". Finally, we gave criteria for recognizing various notions of locality for theory extensions by checking whether certain partial models of the theory extension embed into total models, thus generalizing the results in [2], where links between *locality of sets of equational Horn clauses* and semantical criteria (such as embeddability of partial into total algebras) are established.

We continued this direction of research in several papers. The criteria for detecting the locality of a theory extension established in [11] are used in [13, 5, 6] for giving several examples of local extension. In addition, in [15, 16] we develop a new criterion for detecting locality, based on possibilities of embedding partial models where the extension functions have a *finite definition domain* into total models. This allows us to identify a large number of local theory extensions important in verification and knowledge representation, and to establish in a uniform way upper bounds for the complexity of the universal theory of such extensions. The theory extensions we proved to be local encompass:

(i) *Extensions with free functions, possibly subject to additional boundedness axioms* [3, 4, 13, 15, 16]. Such extensions are important for instance in modeling systems in which some parameters change according to a family of rules which correspond to a partition of the state space [13].

(ii) *Extensions with additional functions + definitions for them (possibly by case distinction)* [13, 15, 16]. We use such extensions in [15] for giving a decision procedure for the universal theory of the class of all MV-algebras.

(iii) *Extensions with constructors and selectors, and with axioms guarded by definedness restrictions* [11, 13]. These results allow us, in particular, to explain from a locality

point of view the results obtained by McPeak and Necula in [7], where they show that local reasoning in pointer data structures is possible.

(iv) *Extensions with (piecewise) monotone functions, possibly subject to boundedness conditions* [13, 15]. Extensions with simple monotonicity conditions were studied already in [11]. In [13] we analyze extensions with (piecewise) monotone functions, possibly subject to boundedness conditions, and identify situations when such extensions are local. The locality of extensions with functions satisfying *generalized monotonicity conditions* and *boundedness* is studied in detail in [15, 16]. We apply these results to automated theorem proving in extensions with monotone functions of posets, as well as of semilattice and lattice-ordered structures. We showed, for instance, that extensions with generalized monotonicity conditions of any theory whose models have a bounded (semi)lattice structure are local, and thus hierarchical reasoning is possible in all such extensions.

Our previous results on parameterized complexity of local theory extensions [11] allow us to estimate the complexity of reasoning in such extensions:

- It allows us to prove, for instance, that if the uniform word problem in a class of algebras with an underlying (semi)lattice structure is decidable in PTIME, the same holds for its extension with monotone functions.
- Another consequence is that the uniform word problem and the universal theory of classes $A(\Sigma, \Sigma')$ of semilattices, lattices, distributive lattices and Boolean algebras with operators in a set Σ satisfying a set $Ax(\Sigma)$ of axioms (e.g. join- and meet-(anti)hemimorphisms) and additional operators in a set Σ' satisfying generalized monotonicity axioms has the same order of complexity as the uniform word problem and the universal theory of the corresponding classes $A(\Sigma)$ of semilattices, lattices, distributive lattices and Boolean algebras with operators in Σ subject to axioms $Ax(\Sigma)$ only.

(v) *Extensions with strictly monotone functions* and applications to checking safety of a systems of trains controlled by a radio control center are studied in [5, 6].

In addition, in [12, 13, 15, 16] we analyze situations when adding comparisons between functions does not destroy the locality of an extension. We also studied combinations of local theory extensions and identified situations when the combination of two local extensions of a base theory is again local [13]. The type of information which needs to be exchanged between the provers for the component (local) theories in this case is described in [12].

We used hierarchical reasoning for automatic invariant checking, i.e. checking whether a given formula is an inductive invariant. An example we considered in the frame of the project AVACS involved the task of proving safety properties for systems of trains on a linear track controlled by a radio control center [5, 6]. By using abstract data types (arrays or monotonely decreasing functions) for storing the train positions we pass in an elegant way from verification of several finite instances of problems to general verification results, in which the number of trains as well as various data of the system (update interval, minimal/maximal allowed speed) are regarded as parameters. By hierarchical reasoning and quantifier elimination we succeeded in generating relations between the parameters of these systems which

guarantee safety. The details are presented in Section 17.7.1. The ideas were further developed in [1], where we extend existing verification methods for CSP-OZ-DC to reason about real-time systems with complex data types and timing parameters. We show that important properties of systems can be encoded in well-behaved logical theories in which hierarchical reasoning is possible. We illustrate the ideas by means of a simplified version of a case study from the European Train Control System standard.

Another area of applications of our results is automated reasoning in non-classical logics. In [15, 16] we used methods for hierarchical reasoning (in extensions with free functions with boundedness and definedness conditions) for giving optimal time decision procedures for the universal theory of MV-algebras, of Gödel algebras and for other similar varieties of algebras. We used hierarchical reasoning in extensions with monotone functions (subject also to certain 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). Such extensions are important in non-classical logics in general and in description logics in particular.

In [9, 14] we give an overview of a variety of results, all centered around a common theme, namely embedding of non-classical logics into first order logic and resolution theorem proving. This study was motivated by the fact that our previous results on automated theorem proving for distributive lattices with operators [10, 8], subsume existing methods for embedding many-valued or modal logics into classical logic. We identify situations when resolution-based decision procedures with optimal complexity can be obtained by using refinements of resolution such as ordered resolution with selection, or ordered chaining with selection, or by using the locality of the equational theories which occur in this context.

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, W. Schulte, and J. S. Dong, eds., *Proceedings of IFM 2007: Integrated Formal Methods*, Oxford, UK, 2007, Lecture Notes in Computer Science. Springer. Accepted for Publication.

[2] H. Ganzinger. Relating semantic and proof-theoretic concepts for polynomial time decidability of uniform word problems. In D. A. Williams, ed., *Proceedings of the 16th IEEE Symposium on Logic in Computer Science (LICS-01)*, Boston, USA, 2001, pp. 81–90. IEEE.

[3] H. Ganzinger, V. Sofronie-Stokkermans, and U. Waldmann. Modular proof systems for partial functions with weak equality. In D. Basin and M. Rusinowitch, eds., *Automated reasoning, Second International Joint Conference, IJCAR 2004*, Cork, Ireland, June 2004, *LNAI 3097*, pp. 168–182. Springer.

- [4] H. Ganzinger, V. Sofronie-Stokkermans, and U. Waldmann. Modular proof systems for partial functions with Evans equality. *Information and Computation*, 204(10):1453–1492, October 2006.

- [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, August 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*, 2007.

[7] S. McPeak and G. Necula. Data structure specifications via local equality axioms. In K. Etessami and S. Rajamani, eds., *Computer Aided Verification, 17th International Conference, CAV 2005*, 2005, *LNCS 3576*, pp. 476–490.

[8] V. Sofronie-Stokkermans. On uniform word problems involving bridging operators on distributive lattices. In U. Egly and C. Fermüller, eds., *Automated Reasoning with Analytic and Related Methods, International Conference, TABLEAUX 2002*, Copenhagen, Denmark, 2002, *LNAI 2381*, pp. 235–250. Springer.

[9] V. Sofronie-Stokkermans. Automated theorem proving by resolution in non-classical logics. In M. Nadif, A. Napoli, E. SanJuan, and A. Sigayret, eds., *Fourth International Conference Journees de l'Informatique Messine: Knowledge Discovery and Discrete Mathematics (JIM-03)*, Metz, France, 2003, pp. 151–167. INRIA. Invited paper.

[10] V. Sofronie-Stokkermans. Resolution-based decision procedures for the universal theory of some classes of distributive lattices with operators. *Journal of Symbolic Computation*, 36(6):891–924, 2003.

• [11] 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.

• [12] 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.

• [13] V. Sofronie-Stokkermans. Local reasoning in verification. In S. Autexier and H. Mantel, eds., *IJCAR'06 Workshop, VERIFY'06: Verification Workshop*, Seattle, USA, August 2006, pp. 128–145.

• [14] V. Sofronie-Stokkermans. Automated theorem proving by resolution in non-classical logics. *Annals of Mathematics and Artificial Intelligence*, 2007. Accepted for Publication.

• [15] V. Sofronie-Stokkermans and C. Ihlemann. Automated reasoning in some local extensions of ordered structures. In *Proceedings of ISMVL 2007*, Oslo, Norway, 2007. IEEE.

• [16] V. Sofronie-Stokkermans and C. Ihlemann. Automated reasoning in some local extensions of ordered structures. *Journal of Multiple-Valued Logic*, p. 18pp, 2007.

17.4.2 Interpolation in local theory extensions

Investigators: Viorica Sofronie-Stokkermans and Andrey Rybalchenko

When verifying the correctness of a system (e.g. correctness of a program, safety of a reactive or hybrid system, or consistency in a complex database) having efficient deduction algorithms is not always sufficient. 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 locating errors. Similarly, 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 with respect to 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. This was the topic of our research described in [8] and [6].

Interpolation in local theory extensions

In [8] 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 [8] 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. [8]);

(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. [8]).

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), and
- verification (e.g. applications to abstraction-refinement and to goal-directed over-approximation (for achieving faster termination)).

The results we obtained in [8] considerably generalize the results on interpolation in extensions of linear arithmetic with free function symbols of McMillan [4].

Interpolation in linear arithmetic with uninterpreted function symbols and applications to program verification

Interpolation [1] is an important component of recent methods for program verification. It provides natural and effective means for computing a separation between sets of 'good' and 'bad' states. Such separations provide a basis for powerful heuristics for the discovery of relevant predicates for predicate abstraction with refinement and for over-approximation in model checking. The applicability of interpolation-based verification methods crucially depends on the employed procedure for interpolant generation. The existing algorithms for interpolant generation which are used for abstraction refinement are proof-based: They require explicit construction of proofs, from which interpolants can be computed (resolution proofs in propositional logic, proofs for linear inequalities over the reals, or in the combined theory of linear arithmetic with uninterpreted function symbols [3, 5, 4]). Explicit construction of such proofs is a difficult task, which hinders the practical applicability of interpolants for verification. In fact, the existing tools for the generation of interpolants over linear arithmetic

and uninterpreted function symbols only handle the difference bound-fragment of arithmetic constraints [2, 4].

In [6] we study possibilities of efficiently computing interpolants with a simple form for extensions of linear arithmetic with free function symbols. We first describe an algorithm for the generation of interpolants for linear arithmetic only, which is based on a reduction to constraint solving. The algorithm has the following advantages:

- it allows to handle directly strict and non-strict inequalities,
- it can be implemented using a Linear Programming solver as a black box.

We then present an algorithm for generating interpolants in combination of linear arithmetic with uninterpreted function symbols, following the hierarchical style of [7, 8]. It applies the algorithm for the generation of interpolants for linear arithmetic as a subroutine. The algorithm presented in [6] differs from that in [8], being tuned to a constrained based approach.

We provide experimental evidence of the applicability of this constraint based interpolant generation. Our implementation is integrated into the predicate discovery procedure of the software verification tools ARMC (http://www.mpi-sb.mpg.de/∼rybal/armc) and Blast (http://embedded.eecs.berkeley.edu/blast/). Our experiments with Blast on Windows device drivers provide a direct comparison with the existing tool FOCI [4], and show promising running times in favor of the constraint based approach. Our method can handle systems which pose problems to other interpolation-based provers: it allows us to handle problems involving both strict and non-strict inequalities, and to verify examples requiring predicates over up to four variables (cf. also Section 17.7.1).

References

[1] W. Craig. Linear reasoning. A new form of the Herbrand-Gentzen theorem. *Journal of Symbolic Logic*, 22(3):250–268, 1957.

[2] R. Jhala and K. L. McMillan. A practical and complete approach to predicate refinement. In *Proceedings of TACAS'2006*, 2006, *LNCS 3920*, pp. 459–473. Springer.

[3] J. Krajícek. Interpolation theorems, lower bounds for proof systems, and independence results for bounded arithmetic. *Journal of Symbolic Logic*, 62(2):457–486, 1997.

[4] K. L. McMillan. An interpolating theorem prover. *Theoretical Computer Science*, 345(1):101–121, 2005.

[5] P. Pudlák. Lower bounds for resolution and cutting plane proofs and monotone computations. *Journal of Symbolic Logic*, 62(3):981–998, 1997.

• [6] A. Rybalchenko and V. Sofronie-Stokkermans. Constraint solving for interpolation. In B. Cook and A. Podelski, eds., *8th International Conference on Verification, Model Checking and Abstract Interpretation (VMCAI 2007)*, Nice, France, 2007, *LNCS 4349*, pp. 346–362. Springer.

• [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. 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.

17.4.3 Modular Verification

Investigator: Viorica Sofronie-Stokkermans

This direction of our work was motivated by the following questions:

- Which properties of complex systems can be checked in a modular way?
- How should the systems be interconnected to make modular verification (e.g. of safety properties) possible?

The answer to such questions depends on how systems and their interaction are modeled. Our work in this direction relies on earlier results [1, 4], in which we used an analogy with phenomena in topology and algebraic geometry, where sheaves are used to describe locally defined objects which can be patched together to a global object.

In [2, 3] we present an overview of previous results in which we showed that many notions needed for expressing properties of systems (states, parallel actions, transitions, behavior, traces of execution, and time) have the property that "local" observations (specific to a subset of individual systems) which agree on the common parts can be patched together to a "global" observation (which refers to the system obtained by interconnecting the individual subsystems) and, thus, they can be modeled as sheaves over a suitable topological space. This allowed us to prove, by using results from geometric logic, that those properties of systems that can be expressed by *cartesian axioms* are preserved after interconnecting the systems [4]. In addition, in [2, 3] possibilities of modeling the behavior of complex systems by means of partially commutative monoids, and links with sheaf representation theorems in algebra are presented.

In [3] we illustrate these ideas by means of an example involving a family of interacting controllers controlling subsets of consecutive trains on a linear, loop-free, rail track. The syntactical description of classes of properties of systems which are guaranteed to be preserved under interconnection [4] allows us to identify examples of safety and lifeness properties of such systems of controllers which can be checked in a modular way.

References

[1] V. Sofronie-Stokkermans. *Fibered Structures and Applications to Automated Theorem Proving in Certain Classes of Finitely-Valued Logics and to Modeling Interacting Systems.* PhD thesis, RISC-Linz, J. Kepler University Linz, 1997.

- [2] 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, August 2006.

- [3] V. Sofronie-Stokkermans. Sheaves and geometric logic and applications to modular verification of complex systems. *Electronic Notes in Theoretical Computer Science*, p. 25 p., 2007. accepted for publication.

[4] 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.

17.4.4 Proof by Consistency

Investigators: Matthias Horbach and Christoph Weidenbach

The goal of inductive theorem proving is to prove (or refute) formulas that are not first-order consequences of a given theory but hold in distinguished models. If, e.g., the natural numbers with addition are axiomatized by the equations $\forall x\ x + 0 \simeq x$ and $\forall x, y\ x + s(y) \simeq s(x + y)$, there are interpretations in which the formula $\forall x\ 0 + x \simeq x$ does not hold. However, it holds in the initial model, as one can easily prove by induction.

As the induction theorem is a higher-order theorem and thus cannot be handled by first-order theorem provers, there are efforts to circumvent direct induction [1]. Around 1980, Lankford [4] developed the proof by consistency technique with which one can attack inductive validity without using an explicit induction scheme. Ganzinger/Stuber [3] and Comon/Nieuwenhuis [2] extended the applicability of the method to a range of specifications: Ganzinger/Stuber found a semi-decision procedure for general saturated sets of universal first-order clauses and Comon/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 perfect model of the clause set if its saturation terminates without generating a contradiction and without changing the perfect model at hand.

To this end, the approach of Comon/Nieuwenhuis requires the a priori computation of so called I-axiomatizations that exclude undesired first-order models and hence, does not require restrictions to saturation inferences. However, it is not known how I-axiomatizations can be computed in general and in particular for non Horn clauses. Ganzinger/Stuber do not require I-axiomatizations at the price of restricted saturation inferences, causing their method to only terminate on a few clause classes.

We want to combine both approaches. The idea is to generate I-axiomatizations during saturation. Whenever a clause is generated that might be in conflict with the perfect model, we separate it into its parts defining the I-axiomatization, first-order counter examples, valid instances, and instances subject to further saturation.

References

[1] H. Comon. Inductionless induction. In A. Robinson and A. Voronkov, eds., *Handbook of Automated Reasoning*, pp. 913–962. Elsevier, 2001.

[2] H. Comon and R. Nieuwenhuis. Induction = I-axiomatization + first-order consistency. *Information and Computation*, 159(1/2):151–186, May 2000.

[3] 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.

[4] D. S. Lankford. A simple explanation of inductionless induction. Memo mtp-14, Louisiana Technical University, Dep. of Math., Ruston, 1981.

17.5 Decision Procedures

17.5.1 Superposition and Decision Procedures

Investigators: Thomas Hillenbrand, Carsten Ihlemann, Uwe Waldmann, and Christoph Weidenbach

In an invited talk at the 2002 Federated Logic Conference, Natarajan Shankar raised a controversy on "big engines vs. little engines" of proof, namely on the question which sort of automated deduction fits more closely to the needs of verification: either full first-order reasoning say with resolution and its refinements, or Nelson-Oppen style combinations of decision procedures for specific domains. Each style comes with its pros and cons: The latter is a push-button technology, but covers only particular theories, and usually is limited to the universal fragment. In contrast, the former offers a richer expressiveness, but in general is a semi-decision procedure only.

One response to the engine debate is to employ big engines as little engines: Resolution-style calculi can be turned into decision procedures for many interesting theories, usually by the formulation of adequate inference and simplification strategies. Consequently the combination with reasoning outside the particular theory comes more or less for free.

From a broader perspective, deduction in the combination of built-in and free theories has been an important topic of research for more than twenty years. Numerous algebraic theories have tightly been integrated into general deductive systems in a white-box fashion, for example Abelian semigroups ([7], among others) or cancellative Abelian monoids [9]. Black-box integration approaches are formulated for whole classes of theories that satisfy particular properties. Examples include theory resolution [8] and superposition modulo a Shostak theory [3].

Our current focus is on application-driven specializations of the superposition calculus, thereby reformulating decidability results in a superposition framework. Initiated by the Verisoft project, we undertook a case study in processor verification, namely verifying the data path of a DLX-like processor with the superposition-based theorem prover SPASS. We managed to prove correctness with SPASS in a fully automatic fashion. This success was possible because of a new encoding for the theory of bitvectors. We have been able to show that, using this new theory presentation, superposition with standard simplifications decides the validity of universal formulae. Hence SPASS just off-the-shelf implements this decision procedure.

Another line of research deals with finite sorts [5]. Reasoning about them in resolution calculi is very often painful. Despite the principal decidability, superposition implementations typically will not terminate. Our main motivation to study this problem is that finite domains often occur in combination with infinite ones. Think for example of finite enumeration types in programming languages, or any verification problem that involves a component with finite state space.

If run without care, standard superposition need not terminate even on the theory axiomatization alone. In case $n = 2$ the finite domain clause $x \simeq 1 \vee x \simeq \underline{2}$ overlaps with itself at the underlined position and produces the resolvent $x \simeq y \vee x \simeq 1 \vee y \simeq 1$. The latter is the

starting point for an infinite sequence of clauses

$$x_1 \simeq 1 \vee x_2 \simeq 1 \vee \left(\bigvee_{i=2}^{2^k-1} x_i \simeq x_{i+1} \right) \vee x_{2^k} \simeq x_1$$

no element of which subsumes any other; and indeed this sequence can be observed for a number of popular theorem provers.

Via exhaustive instantiation, such finite domain problems can be turned into ground satisfiability problems, blowing up the clause set exponentially many in the number of variables. The resulting class is decided by superposition plus splitting, provided inferences are carried out as simplifications. In order to evaluate this in practice, we have encoded correspondingly the currently highly popular Sudoku puzzles [4]. To our surprise, SPASS solves the puzzles within a blink of an eye. This encouraged us to attempt a calculus-level treatment of finite domains, lifting exactly these ground-level inferences.

Very recently, we have observed that lifting in the case of finite domains can be made more economical: A variable needs to stand no longer for any ground term, but just for the finitely many digits that represent the domain. Conversely, inferences involving a most general unifier σ only have to be considered if the range of σ consists of variables and digits. In other words, no complex unifiers are needed; and inferences do not increase the number of variables. This improves upon what is obtained with the basicness restriction [6, 1] in the general case. Secondly, for any non-ground inference we can easily determine those instances that satisfy its ordering constraints. Thirdly, redundancy also needs to refer to digit instances only, such that stronger simplifications become possible.

This lifting modification applies to the family of superposition calculi. In a calculus configuration with a positive unit literal strategy [2] and an aggressive splitting rule, we obtain a decision procedure for satisfiability modulo the finite domain clause. Notably this decides the Bernays-Schönfinkel class as well. Ongoing work is to exploit the benefits of our approach for reasoning in ontologies, maybe even for dealing with cardinalities in a limited way.

References

[1] L. Bachmair, H. Ganzinger, C. Lynch, and W. Snyder. Basic paramodulation. *Information and Computation*, 121(2):172–192, 1995.

[2] N. Dershowitz. A maximal-literal unit strategy for Horn clauses. In S. Kaplan and M. Okada, eds., *Proceedings of the 2nd International Workshop on Conditional and Typed Rewriting Systems*, 1991, *LNCS 516*, pp. 14–25. Springer-Verlag.

[3] H. Ganzinger, T. Hillenbrand, and U. Waldmann. Superposition modulo a Shostak theory. In F. Baader, ed., *Automated Deduction, CADE-19, 19th International Conference on Automated Deduction*, Miami, Florida, July 2003, *LNAI 2741*, pp. 182–196. Springer.

- [4] T. Hillenbrand, D. Topic, and C. Weidenbach. Sudokus as logical puzzles. In H. de Nivelle, ed., *Disproving'06: Non-Theorems, Non-Validity, Non-Provability*, Seattle, USA, August 2006, pp. 2–12. Self publishing.

- [5] T. Hillenbrand and C. Weidenbach. Superposition for finite domains. Research Report MPI-I-2007-RG1-002, Max-Planck Institute for Informatics, Saarbruecken, Germany, April 2007.

[6] R. Nieuwenhuis and A. Rubio. Theorem proving with ordering and equality constrained clauses. *Journal of Symbolic Computation*, 19(4):321–351, 1995.

[7] G. D. Plotkin. Building-in equational theories. In B. Meltzer and D. Michie, eds., *Machine Intelligence 7*, ch. 4, pp. 73–90. American Elsevier, 1972.

[8] M. E. Stickel. Automated deduction by theory resolution. *Journal of Automated Reasoning*, 1(4):333–355, 1985.

[9] U. Waldmann. Cancellative abelian monoids and related structures in refutational theorem proving (Part I). *Journal of Symbolic Computation*, 33(6):777–829, June 2002.

17.5.2 Unification in distributive lattices

Investigator: Viorica Sofronie-Stokkermans

In [7] we give decision procedures for the following problems:

- unification with and without constants in the class D_{01} of bounded distributive lattices,
- unification with linear constant restrictions in D_{01},
- the positive theory of D_{01},
- unification *over* (i.e. in an algebraic extension of) the free distributive lattice with n generators.

The study was motivated, on the one hand, by our interest in giving decision procedures for various fragments of the theory of distributive lattices with operators, and, on the other hand, by the fact that unification problems in semilattice- and lattice-based structures are becoming of increasing interest in computer science (cf. e.g. the results of Baader and Narendran on unification of concept terms in description logics [1]; similar applications in set constraints may also be of interest). The paper extends and improves results presented in [6].

The method we propose uses the fact that the free distributive lattice with n generators can be represented as a lattice of upwards closed subsets (order filters) of a finite partially ordered set. Based on this idea, we propose an algorithm which consists of two steps:

1. *Structure-preserving translation to clause form:* testing the satisfiability of a unification problem $\mathcal{S}$ is reduced to the problem of checking the satisfiability of a set $\Phi_{\mathcal{S}}$ of ground clauses.
2. Testing the satisfiability of $\Phi_{\mathcal{S}}$ can be done using any method for checking satisfiability of sets of clauses in propositional logic.

We used similar ideas for giving a decision procedure for unification with linear constant restrictions. As a byproduct, using Prop. 5.6 in [2], our results show that standard methods for satisfiability checking can be used for deciding the positive theory of D_{01}.

For running tests on several examples we used two different methods, both based on the structure-preserving translation to clause form. The first approach uses first-order logic as much as possible (e.g. we encode subset inclusion in first order logic); SPASS is used to check the satisfiability of the resulting set of clauses. The second approach is purely propositional (we use the fact that reasoning is done in a finite domain). We used FLOTTER to generate an optimized translation to clause form, and checked the satisfiability of the resulting set of clauses with SPASS and zChaff [5].

We also analyzed unification *over* (i.e. in an algebraic extension of) the free lattice. We showed that this problem can be reduced to a satisfiability problem for $\forall\exists$ quantified boolean formulae, and proved that the unification problem with free constants in the class of distributive lattices *over* the free distributive lattice is Π_2^p-complete. We made tests both with SPASS (with a non space-optimal encoding of the problems into first order logic) and with the QBF solvers QuBE [3] and SEMPROP [4].

References

[1] F. Baader and P. Narendran. Unification of concept terms in description logics. In H. Prade, ed., *Proceedings of ECAI'98*, 1998, pp. 331–335. Wiley.

[2] F. Baader and K. Schulz. Unification in the union of disjoint equational theories: Combining decision procedures. *J. Symbolic Computation*, 21:211–243, 1996.

[3] E. Giunchiglia, M. Narizzano, and A. Tacchella. QuBE a system for deciding quantified Boolean formulas satisfiability. In R. Goré, A. Leitsch, and T. Nipkow, eds., *Automated Reasoning, First International Joint Conference, IJCAR 2001, Proceedings*, 2001, LNAI 2083, pp. 364–369. Springer Verlag.

[4] R. Letz. Lemma and model caching in decision procedures for quantified Boolean formulas. In C. Fermüller and U. Egly, eds., *Proceedings of the International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2002)*, 2002, LNAI 2381, pp. 160–175. Springer Verlag.

[5] M. Moskewicz, C. Madigan, Y. Zhao, L. Zhang, and S. Malik. Chaff: Engineering an efficient SAT solver. In *Proceedings of the 39th Design Automation Conference (DAC 2001)*, 2001, pp. 530–535. ACM.

[6] V. Sofronie-Stokkermans. On unification for bounded distributive lattices. In D. McAllester, ed., *Proceedings of the 17th International Conference on Automated Deduction (CADE-17)*, Pittsburgh, Pennsylvania, USA, 2000, *LNAI 1831*, pp. 465–481. Springer.

• [7] V. Sofronie-Stokkermans. On unification for bounded distributive lattices. *ACM Transactions on Computational Logic*, 8(2), April 2007.

17.6 First-Order Model Checking

Investigators: Swen Jacobs and Uwe Waldmann in cooperation with Werner Damm (Univ. Oldenburg), Christoph Scholl (Univ. Freiburg) et al.

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.

Together with the AVACS teams in Freiburg and Oldenburg, we have addressed the analysis of hybrid systems with a focus on the discrete part [2]. Systems with non-trivial discrete state spaces 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 [1], 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. The discrete part of the state is encoded in bit vectors of fixed length. Sets of discrete states are represented in an efficient format for boolean functions, in our case functionally reduced AND-Inverter graphs (FRAIGs) [4]. The state vectors are extended by additional components referring to linear (first-order) constraints. Model checking works essentially as in [1, 5] on the discrete part, while in parallel for the continuous part a Hoare-like calculus is applied.

To make an automatic proof procedure out of this, we add diverse reasoning procedures for the first-order constraints. HySAT [3] is used as an SMT procedure to check whether a fixpoint has been reached during model checking. However, a key point of our approach is the idea to avoid expensive applications of decision procedures as much as possible. Test vector generation for fast inequality checks of boolean combinations of constraints, implication checks for linear constraints, and advanced boolean reasoning are examples for methods which provide some lightweight and inexpensive reasoning and are used both in the context of subsumption checks and for keeping state set representations as compact as possible.

References

[1] J. R. Burch, E. M. Clarke, K. L. McMillan, D. L. Dill, and J. Hwang. Symbolic model checking: 10^{20} states and beyond. In *Proceedings of the 5th Annual IEEE Symposium on Logic in Computer Science*, 1990, pp. 428–439. IEEE press.

- [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] M. Fränzle and C. Herde. Efficient proof engines for bounded model checking of hybrid systems. *ENTCS*, 133:119–137, 2005.

[4] 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.

[5] 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.

17.7 Applications

Taking automation of logics seriously amounts to test our research results in practice. To this end we implement our ideas into model checkers or theorem provers and attack relevant problems. We present three examples of such work in this section. The first one refers to a case study where we could eventually prove safety properties of radio controlled train systems without assuming a concrete number of involved trains. The result is based on deep insights into the combination of logical theories. The second example deals with the automatic analysis of entire IT-infrastructures. The case study gives an insight into the state of the art in first-order theorem proving with respect to manageable first-order fragments, feasible detail of the involved models and eventual performance of the implemented algorithms.

17.7.1 Local Reasoning in Verification

Investigators: Swen Jacobs, Andrey Rybalchenko, and Viorica Sofronie-Stokkermans

We applied the methods for reasoning in local theory extensions and in combinations of theories described in Section 17.4 in various areas, in particular in verification.

Invariant checking, bounded model checking

The simplest area of application is automatic invariant checking, i.e. checking whether a given formula is an inductive invariant. An example we considered in the frame of the project AVACS involved the task of proving safety properties for systems of trains on a linear track controlled by a radio control center [5, 6]. We considered a radio block center (RBC), which is responsible for a given track segment and communicates with all trains that are within this area. Trains may enter and leave 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.

With traditional methods it is difficult to prove correctness of the system as described in full generality: one usually needs to consider various finite instances of the problem.

The approach we use in [5, 6] is different from previously used methods. We use sorted arrays (or monotonely decreasing functions) for storing the train positions. The use of abstract data structures allows 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, and the number of trains is regarded as a parameter. We show that for invariant or bounded model checking the specific properties of "position updates" can be expressed in a natural way by using chains of local theory extensions. Therefore we can use results in hierarchic theorem proving both for invariant and for bounded model checking. By using locality of logical theories we could also obtain formal arguments on possibilities of systematic "slicing" (for bounded model checking). In addition, by using 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 (update interval, minimal/maximal allowed speed) which guarantee safety.

This direction of research is continued in [1]. In the specification of complex systems, one needs to take several aspects into account: control flow, data changes, and timing aspects. Motivated by this necessity, in [4, 2] a specification language CSP-OZ-DC (COD) is defined, which combines Communicating Sequential Processes (CSP), Object-Z (OZ) and the Duration Calculus (DC). Existing verification techniques for COD [3, 8] do not incorporate data structures such as lists and arrays. In [1] we use COD specifications of systems, with complex data types and timing parameters, and analyze possibilities for efficient invariant checking and bounded model checking in these systems. The main contributions of the paper can be described as follows.

Specification: We extend work in which COD specifications were used [4, 2, 8] in two ways:

(i) We use abstract data structures (arrays, functions) for representing and storing information about an unspecified *parametric* number of components of the systems. The use of abstract data structures allows us to pass in an elegant way from verification of several finite instances of a verification problem (for 2, 3, 4, ... components) to general verification results, in which the number of components is a parameter.

(ii) In order to achieve a tighter binding of the OZ to the DC part, we introduce timing parameters. This enables more flexible specifications of timing constraints and, additionally, for referring to time constants also within the specification's data part.

Verification: We show that, in this context, invariant checking or bounded model checking can be reduced to proving in complex theories. We analyze the theories that occur in relationship with a given COD specification, and present a sound method for efficient reasoning in such theories. We also identify situations when the method is sound and complete (i.e., when the specific properties of systems define chains of local theory extensions).

Applications: Our running example is an extension of a case study that we considered in [5, 6]. Here, we additionally encompass efficient handling of emergency messages and illustrate the full procedure – starting from a COD description of the case study to verification.

Verification by abstraction-refinement

The algorithm for hierarchical computation of interpolants described in [10] 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 Viorica Sofronie-Stokkermans with Andrey Rybalchenko [9] (cf. also Section 17.4.2). Our 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), and thus can be used for abstraction-based verification in examples involving extensions of linear arithmetic with free function symbols. Our method can handle systems which pose problems to other interpolation-based provers: it allows us to handle

problems involving both strict and non-strict inequalities, and to verify examples requiring predicates over up to four variables. Several experiments with Blast on Windows device drivers provide a direct comparison with the existing tool FOCI [7], and show promising running times in favor of the constraint based approach.

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, W. Schulte, and J. S. Dong, eds., *Proceedings of IFM 2007: Integrated Formal Methods*, Oxford, UK, 2007, Lecture Notes in Computer Science. Springer. Accepted for Publication.

[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. Fitzgerald, I. Hayes, and A. Tarlecki, eds., *FM 2005*, 2005, *LNCS 3582*. Springer.

[4] J. Hoenicke and E.-R. Olderog. CSP-OZ-DC: A combination of specification techniques for processes, data and time. *Nordic Journal of Computing*, 9(4):301–334, 2002. appeared March 2003.

- [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, August 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*, 2007.

[7] K. L. McMillan. An interpolating theorem prover. *Theoretical Computer Science*, 345(1):101–121, 2005.

[8] R. Meyer, J. Faber, and A. Rybalchenko. Model checking duration calculus: A practical approach. In *ICTAC*, 2006, *LNCS 4281*, pp. 332–346. Springer.

- [9] A. Rybalchenko and V. Sofronie-Stokkermans. Constraint solving for interpolation. In B. Cook and A. Podelski, eds., *8th International Conference on Verification, Model Checking and Abstract Interpretation (VMCAI 2007)*, Nice, France, 2007, *LNCS 4349*, pp. 346–362. Springer.

- [10] 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.

17.7.2 Automatic Analysis of IT-Infrastructures

Investigators: Simon Hirth, Carsten Karl, and Christoph Weidenbach

The configuration, maintenance and debugging of IT-infrastructures is a non-trivial task. It ends up in harmonizing the individual configurations of network components such as routers or firewalls, services provided by servers such as DHCP, http or mail, and clients eventually consuming the services. Furthermore, any attempt in setting up an IT-infrastructure must fulfill several conflicting goals such as security, performance, robustness or maintainability, typically performed by different (groups of) people, resulting in a serious error potential.

Our approach is to analyze infrastructures with respect to defects, e.g., bugs in the configuration that prevent a client from obtaining a service. Starting from our work on security

protocols [5], we developed an integrated model that starts at ethernet level and can handle all IT-infrastructure functionality up to higher level protocols such as DHCP [3].

The ethernet model is a set of sent messages, where packets are represented by first-order terms. Time is modelled by sequences of messages and all communication channels are assumed to be error free. For example, the atom

$$Send(epacket(ethernet1, dstmac, srcmac, ip, \\ ippacket(srcip, dstip, type, payload)))$$

represents a sent IP-packet on the ethernet *ethernet1*.

A router [2] or firewall [1] is represented by its forwarding behavior. For example, the (simplified) formula

$$[RouteIP(packet(xsrc, xdst, xpld)), RouteEntry(xmsk, xnet, xhop), \\ ipand(xdst, xmsk) \simeq xnet] \rightarrow Send(xhop, packet(xsrc, xdst, xpld))$$

models the forwarding behavior of a router with respect to its routing table, represented by the predicate *RouteEntry*,where every symbol starting with an x represents a universally quantified variable. IP-addresses are represented by 32-bit vectors and the logical and on bitvectors (*ipand*), needed to determine network or interface matches, is modeled by a conditional rewrite system.

Eventually, protocols such as DHCP are formalized via I/O-automata that are eventually translated into first-order logic formulas. The overall theory has a size of about 150 kB of first-order formulas. The work is a balance of (i) using only first-order fragments that can be decided by superposition (SPASS) in reasonable time (ii) a natural and modular and therefore extendable formalization of the involved concepts and (iii) a model that provides maximum precision and detail.

We succeeded in saturating the overall infrastructure configuration of both MPIs in about 4 hours. Then we were, e.g., able to prove that a client connected to an ethernet segment of the institute receives all its services [4].

References

[1] Cisco Systems, Inc. . Catalyst 6500 6500 series switch and cisco 7600 series router firewall services module configuration guide release 3.1(1), 2006.

[2] F. Baker. RFC 1812: Requirements for IP version 4 routers, June 1995. Obsoletes RFC1716, RFC1009. Status: PROPOSED STANDARD.

[3] R. Droms. RFC 2131: Dynamic host configuration protocol, Mar. 1997. Obsoletes RFC1541. Status: DRAFT STANDARD.

- [4] 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.

[5] 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.

17.7.3 Labelled Clauses

Investigators: Christoph Weidenbach in collaboration with Tal Lev-Ami, Thomas Reps, and Mooly Sagiv

Labelled deductive systems (see e.g., [1]) have been pushed forward by Dov Gabbay in the last fifteen years and have become recognized as a significant component of the logic culture, in particular in the context of non-classical logics. We add labels to first-order clauses to simultaneously apply superposition to several proof obligations inside one clause set. From a theoretical perspective, the approach unifies a variety of deduction modes. These include different strategies such as set of support, as well as explicit case analysis, e.g., splitting. From a practical perspective, labelled clauses offer advantages in the case of related proof obligations resulting from multiple conjectures over the same axiom set or from a single conjecture that is a large conjunction. Here we can share clauses (e.g., the axioms and clauses deduced from them, share Skolem symbols), share deduced clause variants, and transfer lemmas between the different obligations. Motivated by software verification, we have created a prototype implementation of labelled clauses that supports multiple conjectures, and we provide convincing experiments for the benefits [2].

References

[1] D. Basin, M. D'Agostino, D. M. Gabbay, S. Matthews, and L. Viganò, eds. *Labelled Deduction*, *Applied Logic Series*, vol. 17. Kluwer, 2000.

- [2] T. Lev-Ami, C. Weidenbach, T. Reps, and S. Mooly. Labelled clauses. In *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007. Springer. Accepted for Publication.

17.8 Software

17.9 Software Systems

The development of automated reasoning systems is a key topic of our group. On the one hand, we want to demonstrate that our theoretical concepts are practically useful. On the other hand, the evolution of concepts is often fertilized when some workbench allows to easily manipulate instances thereof. We provide first-class automated theorem provers which are used world-wide in research groups and industrial contexts.

17.9.1 SPASS

Investigators: Thomas Hillenbrand and Christoph Weidenbach in collaboration with Renate Schmidt, Dalibor Topic, and Rostislav Rusev

SPASS is an automated theorem prover for full first-order logic with equality and a number of non-classical logics. We use SPASS as a platform to evaluate our theoretical results starting from well-chosen small experiments up to real world case studies.

New in the freshly released version SPASS 3.0 [2] 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 play an important role e.g. in the specification of multi-agent systems. *EML* logics also include dynamic modal logics [1] 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 or 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 into first-order logic.

For most decidable *EML* logics, SPASS is actually a decision procedure on the first-order formulas resulting from the translation. For some logics, e.g., description logics including negation of roles, it is currently the only available decision procedure. SPASS is competitive even with special-purpose systems.

Furthermore, SPASS 3.0 offers additional renaming and selection strategies, in particular needed for our case study on the analysis of IT-Infrastructures (see Section 17.7.2), and an improved user/machine interface including an extended formula-clause relationship handling, input and output of saturated clause sets and documentation.

SPASS 3.0 is available from its homepage at `http://spass.mpi-inf.mpg.de/` and offers source as well as binary distributions for Unix, Linux, MacOS and Windows.

References

[1] R. A. Schmidt and U. Hustadt. First-order resolution methods for modal logics. In A. Podelski, A. Voronkov, and R. Wilhelm, eds., *Volume in memoriam of Harald Ganzinger*, LNCS. Springer, 2006. Invited overview paper, to appear.

- [2] C. Weidenbach, R. Schmidt, T. Hillenbrand, R. Rusev, and D. Topic. Spass version 3.0. In *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007. Springer. Accepted for Publication.

17.9.2 SPASS+T

Investigators: Virgile Prevosto and Uwe Waldmann

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 [3]) adds arithmetic reasoning capabilities to SPASS in three complementary ways: First, it uses a set of standard axioms such as $\forall x.\,(x + 0 \simeq x)$ or $\forall x, y.\,((x - y) + y \simeq x)$. Second, built-in simplification rules are employed to reduce numeric subexpressions and to find solutions for variables, so that, for instance, a clause $\forall x.\,(\neg\, x + 2 \simeq 5 + 1 \vee p(x))$ is replaced by $\forall x.\,(\neg\, x \simeq 4 \vee p(x))$ 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 CVC Lite (Barrett and Berezin [1]) or Yices (Dutertre and De Moura [2]). In this combination, SPASS uses its deduction rules to generate formulas as usual; in addition, it passes to the SMT procedure all the formulas that can be handled by the procedure, i. e., all ground formulas. As soon as one of the two systems encounters a contradiction, the problem is solved. It is important to notice that such an

integration of first-order theorem provers and SMT procedures can only be a pragmatic one: Completeness for first-order logic plus arithmetic can be achieved for special classes of inputs, but not in general.

References

[1] C. Barrett and S. Berezin. CVC Lite: A new implementation of the Cooperating Validity Checker. In R. Alur and D. A. Peled, eds., *Computer Aided Verification, 16th International Conference, CAV 2004*, Boston, MA, USA, 2004, *LNCS 3114*, pp. 515–518. Springer-Verlag.

[2] 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.

- [3] 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.

17.9.3 Waldmeister

Investigator: Thomas Hillenbrand in collaboration with Bernd Löchner (Technische Universität Kaiserslautern)

Tailored for reasoning in varieties, WALDMEISTER [2] implements unfailing completion with refinements towards ordered completion (cf. [1]). The system saturates the input axiomatization, distinguishing active facts, which induce a rewrite relation, and passive facts, which are the one-step inference conclusions of the active ones up to redundancy. The saturation process is parameterized by a reduction ordering and a heuristic assessment of passive facts. These parameters are instantiated according to the algebraic structure that is present in the problem at hand.

As in previous years, WALDMEISTER has continued to dominate, in its category, the CADE ATP System Competition [3], which is "the world championship for 1st order automated theorem proving". An interesting new feature is a cancellation rule, which is activated in cancellative domains. In essence, equations $s + t \simeq s + t'$ and $t + s \simeq t' + s$ are simplified into $t \simeq t'$. This simplification takes place on the literal level, as opposed to rewriting which operates on the term level. Though not required for completeness, it turns out that relevant passive facts are selected earlier such that some proofs are found faster.

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. Citius altius fortius: Lessons learned from the theorem prover WALDMEISTER. In I. Dahn and L. Vigneron, eds., *Proceedings of the 4th International Workshop on First Order Theorem Proving, FTP'03*, Valencia, Spain, June 2003, *Electronic Notes in Theoretical Computer Science*, vol. 86.1, pp. 1–13. Elsevier.

[3] G. Sutcliffe and C. Suttner. The state of CASC. *AI Communications*, 19(1):35–48, 2006. Competition archive available via `http://www.tptp.org/CASC`.

17.10 Academic Activities

17.10.1 Conference and Workshop Positions

Membership in Program Committees

Christoph Weidenbach:

- *Tenth International Conference on Foundations of Software Science and Computation Structures, FoSSaCS 2007*, Braga, Portugal, March/April 2007.

Thomas Hillenbrand:

- *6th International Workshop on the Implementation of Logics*, Phnom Penh, Cambodia, November 2006.

Viorica Sofronie-Stokkermans:

- *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2005)*, Koblenz, September 2005,
- *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2007)*, Aix en Provence, July 2007,
- *Automated Deduction: Decidability, Complexity, Tractability (ADDCT'07), CADE-21 Workshop*, Bremen, July 2007.

Uwe Waldmann:

- *ESCoR, FLoC'06 Workshop on Empirically Successful Computerized Reasoning*, Seattle, August 2006,
- *CADE-21 Workshop on Empirically Successful Automated Reasoning in Large Theories*, Bremen, July 2007.

Membership in Organizing Committees

Christoph Weidenbach:

- 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.

Viorica Sofronie-Stokkermans:

- *Automated Deduction: Decidability, Complexity, Tractability (ADDCT'07), CADE-21 Workshop*, Bremen, July 2007.
- *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2007)*, Aix en Provence, July 2007 (Workshop Chair).

17.10.2 Invited Talks and Tutorials

Christoph Weidenbach:

- *From (Security) Protocol to Enterprise Network Infrastructure (Security) Analysis*, Invited talk, Workshop "Trustworthy Software", Saarbrücken, May 2006.
- *Mechanising First-Order Logic: Technology, Decidability and Applications*, Invited talk, University of Cambridge, Computer Science Department.

Viorica Sofronie-Stokkermans:

- *Entscheidungsverfahren für Erweiterungen und Kombinationen von Theorien und Anwendungen*, Invited talk, University of Trier, June 2005.
- *Automatisches Beweisen in Algebra und Anwendungen in der Informatik*, Invited talk, T.U. Dresden, July 2005.
- *Algebraic and logical methods in computer science*, Invited talk, University of Leicester, August 2005.
- *Hierarchical and modular reasoning in extensions and combinations of theories*, Invited talk, University of Manchester, June 2006.
- *Hierarchical and modular reasoning in extensions and combinations of theories*, Invited talk, University of Liverpool, August 2006.
- *Symbolic Computation in komplexen Theorien*, Invited talk, University Passau, January 2007.
- *Modularität in der Verifikation komplexer Systeme*, Invited talk, University Johann Wolfgang Goethe, Frankfurt am Main, February 2007.
- *On interpolation in local theory extensions*, Invited talk, Seminar "Decision procedures, design and combination", LORIA and INRIA Lorraine, Nancy, February 2007.
- *Hierarchical and modular reasoning in complex theories*, Invited talk, Università degli Studi di Verona, Computer Science Colloquium, May 2007.

Uwe Waldmann:

- *Modular Proof Systems for Partial Functions with Evans Equality*, Invited talk, Heriot-Watt University, Edinburgh, Computer Science Department, July 2005.

17.11 Teaching Activities

Summer Semester 2005

Courses:

LAN Design in Practice (C. Weidenbach)

Winter Semester 2005/2006

Seminars:
Decision Procedures for Logical Theories (V. Sofronie-Stokkermans, U. Waldmann).

Summer Semester 2006

Courses:
Automated Reasoning (U. Waldmann, C. Weidenbach)

Diploma Theses, Bachelor Theses, Master Theses

- Jan Bankstahl: Netflow analysis of SAP R/3 traffic in an enterprise environment, Master thesis (Supervisor: C. Weidenbach).
- Andrea Bastuck: Maschinell unterstützte Analyse eines Sicherheitsprotokolls, Diploma thesis (Supervisor: U. Waldmann).
- Simon Hirth: Automatische Analyse von DHCP und höheren Infrastrukturdiensten mit SPASS, Master thesis (Supervisor: C. Weidenbach).
- Carsten Karl: Automatische Analyse von Layer 3 Netzwerken mit SPASS, Master thesis (Supervisor: C. Weidenbach).
- Michel Ludwig: Extensions of the Knuth-Bendix Ordering with LPO-like Properties, Diploma thesis (Supervisor: U. Waldmann).
- Dalibor Topic: Coding Sudoku Puzzles in Logic, Bachelor thesis (Supervisors: C. Weidenbach, T. Hillenbrand).

17.12 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.
- MPI principal investigators: Harald Ganzinger († June 2004), Andreas Podelski (now in Freiburg), Stefan Ratschan (from January 2005, now in Prague), Viorica Sofronie-Stokkermans (since January 2005), Uwe Waldmann (since January 2005).
- Partners: Albert-Ludwigs-Universität Freiburg, Carl von Ossietzky Universität Oldenburg, Universität des Saarlandes.

Verisoft

Verisoft is a long-term research project funded by the German Federal Ministry of Education and Research. The main goal of the project is the pervasive formal verification of computer systems. The correct functionality of systems, as they are applied, for example, in automotive engineering, in security technology and in the sector of medical technology, are to be mathematically proved.

The proofs are computer aided in order to prevent human error conducted by the scientists involved. The knowledge and progress obtained are expected to assist German enterprise in achieving a stable, internationally competitive position in the professional spheres mentioned above.

- Starting date: July 2003.
- Duration: 8 years.
- Funding: BMBF.
- MPI principal investigators: Harald Ganzinger († June 2004), Andreas Podelski (now in Freiburg), Uwe Waldmann (from July 2005 to June 2006).
- Partners: AbsInt, BMW, DFKI, Infineon, OFFIS, OneSpin Solutions, T-Systems, Saarland University, TU Darmstadt, University Koblenz-Landau, TU München.

17.13 Publications

Journal articles

[1] H. Ganzinger, V. Sofronie-Stokkermans, and U. Waldmann. Modular proof systems for partial functions with Evans equality. *Information and Computation*, 204(10):1453–1492, October 2006.

[2] S. Jacobs and V. Sofronie-Stokkermans. Applications of hierarchical reasoning in the verification of complex systems. *Electronic Notes in Theoretical Computer Science*, 2007.

[3] S. Jacobs and U. Waldmann. Comparing instance generation methods for automated reasoning. *Journal of Automated Reasoning*, 38:57–78, April 2007.

[4] V. Sofronie-Stokkermans. Automated theorem proving by resolution in non-classical logics. *Annals of Mathematics and Artificial Intelligence*, 2007. Accepted for Publication.

[5] V. Sofronie-Stokkermans. On unification for bounded distributive lattices. *ACM Transactions on Computational Logic*, 8(2), April 2007.

[6] V. Sofronie-Stokkermans. Sheaves and geometric logic and applications to modular verification of complex systems. *Electronic Notes in Theoretical Computer Science*, p. 25 p., 2007. accepted for publication.

[7] V. Sofronie-Stokkermans and C. Ihlemann. Automated reasoning in some local extensions of ordered structures. *Journal of Multiple-Valued Logic*, p. 18pp, 2007.

Conference articles

[1] 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.

[2] J. Faber, S. Jacobs, and V. Sofronie-Stokkermans. Verifying CSP-OZ-DC specifications with complex data types and timing parameters. In J. Davies, W. Schulte, and J. S. Dong, eds., *Proceedings of IFM 2007: Integrated Formal Methods*, Oxford, UK, 2007, Lecture Notes in Computer Science. Springer. Accepted for Publication.

[3] T. Hillenbrand, D. Topic, and C. Weidenbach. Sudokus as logical puzzles. In H. de Nivelle, ed., *Disproving'06: Non-Theorems, Non-Validity, Non-Provability*, Seattle, USA, August 2006, pp. 2–12. Self publishing.

[4] 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, August 2006, pp. 15–26.

[5] S. Jacobs and U. Waldmann. Comparing instance generation methods for automated reasoning. In B. Beckert, ed., *Automated Reasoning with Analytic Tableaux and Related Methods, International Conference, TABLEAUX 2005*, Koblenz, Germany, 2005, *LNCS 3702*, pp. 153–168. Springer.

[6] T. Lev-Ami, C. Weidenbach, T. Reps, and S. Mooly. Labelled clauses. In *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007. Springer. Accepted for Publication.

[7] 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.

[8] A. Rybalchenko and V. Sofronie-Stokkermans. Constraint solving for interpolation. In B. Cook and A. Podelski, eds., *8th International Conference on Verification, Model Checking and Abstract Interpretation (VMCAI 2007)*, Nice, France, 2007, *LNCS 4349*, pp. 346–362. Springer.

[9] 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.

[10] 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.

[11] V. Sofronie-Stokkermans. Local reasoning in verification. In S. Autexier and H. Mantel, eds., *IJCAR'06 Workshop, VERIFY'06: Verification Workshop*, Seattle, USA, August 2006, pp. 128–145.

[12] 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, August 2006.

[13] V. Sofronie-Stokkermans and C. Ihlemann. Automated reasoning in some local extensions of ordered structures. In *Proceedings of ISMVL 2007*, Oslo, Norway, 2007. IEEE.

[14] C. Weidenbach, R. Schmidt, T. Hillenbrand, R. Rusev, and D. Topic. Spass version 3.0. In *21st International Conference on Automated Deduction (CADE-21)*, Bremen, Germany, 2007. Springer. Accepted for Publication.

Theses

[1] J. Bankstahl. Netflow analysis of SAP R/3 traffic in an enterprise. Masters thesis, Universität des Saarlandes, June 2006.

[2] A. Bastuck. Maschinell unterstützte Analyse eines Sicherheitsprotokolls. Diploma thesis, Universität des Saarlandes, September 2006.

[3] S. Hirth. Automatische Analyse von DHCP und höheren Infrasturkturdiensten mit SPASS. Masters thesis, Universität des Saarlandes, July 2006.

[4] C. Karl. Automatische Analyse von Layer 3 Netzwerken mit SPASS. Masters thesis, Universität des Saarlandes, July 2006.

[5] M. Ludwig. Extensions of the Knuth-Bendix ordering with LPO-like properties. Masters thesis, Universität des Saarlandes, July 2006.

Technical reports

[1] T. Hillenbrand and C. Weidenbach. Superposition for finite domains. Research Report MPI-I-2007-RG1-002, Max-Planck Institute for Informatics, Saarbruecken, Germany, April 2007.

[2] 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.

18 The Machine Learning Group (RG2)

18.1 Personnel

Head of the Group
Prof. Dr. Tobias Scheffer

Secretary
Ellen Fries

PhD Students
Steffen Bickel
Ulf Brefeld
Michael Brückner
Uwe Dick
Laura Dietz
Isabel Drost (–April 2007)
Peter Haider
Sascha Schulz

18.2 Group Organization

The group meets weekly for the "Machine Learning Journal Club" on Mondays, 3 p.m. During this meeting, either external researchers or group members give talks on their current work and discuss relevant papers. In addition, the group meets biweekly for joint reading groups on machine learning and on link analysis with Department 5.

18.2.1 Structural Learning

Investigators: Ulf Brefeld, Uwe Dick, and Peter Haider

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 [6], support vector machines for structured output spaces [7] and other discriminative learners exploit this principle.

Semi-supervised 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. The multi-view approach to semi-supervised learning is based on the observation that the rate of disagreement of two independent experts upper-bounds the individual error rate of either experts. Implementing this principle for semi-supervised learning on structured variables leads to an algorithm that minimizes the error rate for labeled and the disagreement between two initially independent hypotheses for unlabeled data. Work on semi-supervised learning for sequential data [1] has been distinguished by the Best Paper Award of the European Conference on Machine Learning. Follow-up work on learning tree-structured output variables has led to effective parsers for natural language [2].

A current line of research investigates transductive kernel machines for structured output variables [8]. 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 [3]. The discrete optimization parameters are eliminated and the resulting differentiable problems can be optimized more efficiently.

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 [4], 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 [5].

References

[1] U. Brefeld, C. Büscher, and T. Scheffer. Multi-view discriminative sequential learning. In *Proceedings of the European Conference on Machine Learning*, 2005.

[2] U. Brefeld and T. Scheffer. Semi-supervised learning for structured output variables. In *Proceedings of the International Conference on Machine Learning*, 2006.

[3] O. Chapelle. Training a support vector machine in the primal. Technical Report 147, Max Planck Institute Tübingen, 2006.

[4] T. Finley and T. Joachims. Supervised clustering with support vector machines. In *Proceedings of the Twenty-Second International Conference on Machine Learning*, 2005.

• [5] P. Haider, U. Brefeld, and T. Scheffer. Supervised clustering of streaming data for email batch detection. In *Proceedings of the International Conference on Machine Learning*, 2007.

[6] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems*, 2004.

[7] 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.

• [8] A. Zien, U. Brefeld, and T. Scheffer. Transductive support vector learning for structured variables, 2007. Unpublished Manuscript.

18.2.2 Covariate Shift and 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.

Knowledge transfer between multiple related learning problems is also referred to as *multi-task learning*. Nonparametric hierarchical Bayesian theory provides powerful methods to model dependencies between related tasks. Tresp and Yu [4] and Xue et al. [5] describe hierarchical Bayesian models for classification on multiple tasks using Dirichlet process priors.

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. Known work on learning under covariate shift has addressed the model-based case: Model-based optimization criteria such as the joint likelihood of input and output additionally assess how well the classifier models the distribution of input values. This amounts to adding a term to the criterion that is irrelevant for the task at hand.

Work in this line of research has contributed a discriminative model for learning under arbitrarily different training and test distributions [1]. 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.

Learning under Multiple Test Distributions

In server-sided spam filtering, filters for multiple users are required to perform well. Each user receives messages according to an individual, unknown distribution, reflected only in the unlabeled inbox. The spam filter for a user is required to perform well with respect to this distribution. Labeled messages from publicly available sources can be utilized, but they are governed by a distinct distribution, not adequately representing most inboxes. We devise a method that minimizes a loss function with respect to a user's personal distribution based on the available biased sample. A nonparametric hierarchical Bayesian model furthermore generalizes across users by learning a common prior which is imposed on new email accounts [2].

In conjunction with the ECML-PKDD conference, Steffen Bickel organized the Discovery Challenge 2006 [3]. The challenge provides benchmark data and evaluation protocols for learning from biased data and transfer learning in the email spam filtering application.

References

- [1] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the International Conference on Machine Learning*, 2007.
- [2] 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*, Cambridge, USA, 2007, vol. 19. MIT Press.

[3] S. Bickel, V. Tresp, and M. Seeger, eds. *Proceedings of the ECML-PKDD Discovery Challenge Workshop*, 2006.

[4] V. Tresp and K. Yu. An introduction to nonparametric hierarchical Bayesian modelling with a focus on multi-agent learning. In *Switching and Learning in Feedback Systems*, *LNCS 3355*, pp. 290–312. Springer, 2004.

[5] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.

18.2.3 Learning for Information Retrieval

Investigators: Steffen Bickel, Laura Dietz, Isabel Drost, and Sascha Schulz

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.

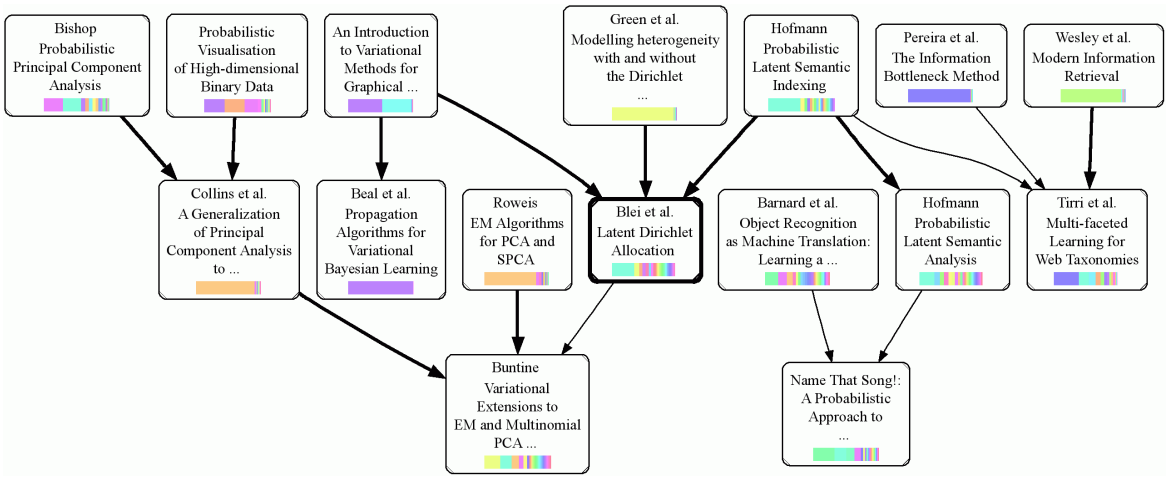


Figure 18.1: Influence strengths of links in the citation vicinity of the publication "Latent Dirichlet Allocation". Strength of arcs represents the strength of influence; colors indicate topics.

Modeling Citation Influences

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 18.1) constitutes a step toward this goal [5, 7, 6]. A probabilistic, generative model 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.

Personalized Ranking

Designing algorithms that rank information items according to their relevance is a key challenge for information retrieval. Exploiting information about a user's specific personal concept of relevance offers an interesting angle at the problem. Such personal information can include a user's click stream or rankings provided by a user. We study how machine learning algorithms can exploit such data sources in order to provide better rankings for individual users. In cooperation with Department 5, we study this problem for search engine queries and online movie recommendation. In cooperation with nugg.ad AG, we investigate this

problem in the context of delivery of online advertisements.

Data and Text Mining in Quality and Service

Workshop reports and warranty databases in the automotive industry contain an abundance of information about quality issues and their latent causes. Probabilistic topic models can again help to identify the latent topics that underly the various textual manifestations of specific symptoms and causes. Probabilistic models can help to discover new trends and to identify factors that are linked to problems whose causes are not yet understood. We investigate in this area in a cooperation with DaimlerChrysler.

User Modeling for Text Input Assistance

Prediction of user behavior has been studied in many application domains, including prediction of the next link that a user will follow and of the next unix shell command that a user will enter [10, 4]. The development of small input devices such as PDAs has fueled work on predicting text input with the goal of providing user assistance [3, 9]. The problem of predicting the succeeding words of an initial fragment of natural language text adds a new dimension to this line of work, because it involves the decoding problem of searching a large space of possible word combinations.

Work on efficient indexing techniques [8], and empirical studies on the effectiveness of language models for text prediction has contributed to this field [1, 2].

References

[1] S. Bickel and T. Scheffer. Learning to complete sentences. In *Proceedings of the European Conference on Machine Learning0*, 2005.

[2] S. Bickel and T. Scheffer. Predicting sentences using n-gram language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2005.

[3] J. Darragh and I. Witten. *The Reactive Keyboard*. Cambridge University Press, 1992.

[4] B. Dasison and H. Hirsh. Predicting sequences of user actions. In *AAAI/ICML Workshop on Predicting the Future: AI Approaches to Time Series Analysis*, 1998.

[5] L. Dietz. Probabilistic topic models to support a scientific community. In *Twenty-First National Conference on Artificial Intelligence*, 2006.

- [6] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the International Conference on Machine Learning*, 2007.

[7] L. Dietz and A. Stewart. Utilize probabilistic topic models to enrich knowledge bases. In *Proceedings of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, 2006.

[8] K. Grabski and T. Scheffer. Sentence completion. In *Proceedings of the SIGIR International Conference on Information Retrieval*, 2003.

[9] C. Kushler. AAC using a reduced keyboard. In *Proceedings of the Conference on Technology and Persons with Disabilities*, 1998.

[10] H. Motoda and K. Yoshida. Machine learning techniques to make computers easier to use. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1997.

18.2.4 Adversarial Learning and 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 [3, 5].

Our investigations in this area are conducted in cooperation with the European web hosting company STRATO AG. Recent work has led to tools that protect STRATO's servers and customers against spam, phishing, and virus emails [2]. Software that will protect STRATO against intrusions and exploitation of their servers for dissemination of spam is currently under development.

Security application have also motivated parts of our work on personalized information filtering [1], and detection of batches [4], discussed in more detail in Section 18.2.1.

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*, Cambridge, USA, 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] 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.

- [4] P. Haider, U. Brefeld, and T. Scheffer. Supervised clustering of streaming data for email batch detection. In *Proceedings of the International Conference on Machine Learning*, 2007.

[5] D. Lowd and C. Meek. Adversarial learning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

18.2.5 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, Lenka Ivantysynova, Humboldt-Universität zu Berlin, and Stefan Wrobel, Fraunhofer IAIS, we investigate how algorithms can be constructed 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. 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. Accurate schema matching on streams. *Intelligent Data Analysis*, 2007. In print.

[2] S. Jaroszewicz and T. Scheffer. Fast discovery of unexpected patterns in data, relative to a bayesian network. In *Proceedings of the ACM SIGKDD Conference*, 2005.

[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.

18.3 Academic Activities

18.3.1 Journal Positions

In 2007, Tobias Scheffer serves on the Editorial Board of the

- *Data Mining and Knowledge Discovery Journal.*

18.3.2 Conference and Workshop Positions

Membership in Steering Committee

In 2007, 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

Since January 2007, members of the group serve on the following Program Committees.

Steffen Bickel

- *European Conference on Machine Learning.* Warsaw, Poland. 2007.
- *European Conference on Principles and Practice of Knowledge Discovery in Databases.* Warsaw, Poland. 2007.
- *International Conference on Machine Learning.* 2007.
- *Pacific-Asian Conference on Knowledge Discovery in Databases.* China. 2007.

Ulf Brefeld

- *European Conference on Machine Learning.* Warsaw, Poland. 2007.
- *European Conference on Principles and Practice of Knowledge Discovery in Databases.* Warsaw, Poland. 2007.

– *ICML Workshop on Constrained Optimization and Structured Output Spaces*. 2007.

Laura Dietz

– *ESCW Workshop on Bridging the Gap between Semantic Web and Web 2.0*. 2007.

Tobias Scheffer

– *International Conference on Machine Learning*. 2007.
– *European Conference on Machine Learning* (Area Chair), 2007.
– *European Conference on Principles and Practice of Knowledge Discovery in Databases* (Area Chair). 2007.

18.3.3 Reviewing for Funding Organizations

In 2007, Tobias Scheffer serves as reviewer of grant proposals of the following organizations.

– *German Science Foundation DFG.*
– *Czech Science Foundation.*
– *Belgian Science Foundation.*

18.3.4 Invited Talks and Tutorials

Tobias Scheffer (since January 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.
– *Maschinelles Lernen und Information Retrieval*. Invited Talk. Technische Universität Chemnitz. 2007.

18.4 Teaching Activities

Winter Term 2006/2007

Courses at Humboldt-Universität zu Berlin:
Maschinelles Lernen und Data Mining (Tobias Scheffer, Steffen Bickel).

Seminars:
Network Mining (Ulf Brefeld, Steffen Bickel, Michael Brückner, Uwe Dick, Laura Dietz, Isabel Drost, Peter Haider, Tobias Scheffer).

18.5 Dissertations, Habilitations, Offers, Awards

18.5.1 Ongoing Dissertation Projects

- Steffen Bickel: Learning for differing training and test distributions.
- Ulf Brefeld: Semi-supervised learning for structured variables.
- Michael Brückner: Adversarial Classification.
- Uwe Dick: Machine learning for intrusion detection.
- Laura Dietz: Probabilistic topic models for the scientific community.
- Isabel Drost: Machine learning for personalized ranking.
- Peter Haider: Structural Learning.
- Sascha Schulz: Improving quality assurance processes in the automotive industry with machine learning.
- David Vogel: Data analysis for insurance risk assessment.

18.6 Grants and Cooperations

18.6.1 Personalized Ranking of Online Avertisements

Funding: nugg.ad AG
Duration: since January 2007.
Project Members: Steffen Bickel, Peter Haider

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.

18.6.2 Data and Text Mining in Quality and Service

Funding: DaimlerChrysler AG
Duration: 08/2005-09/2008
Project Members: Sascha Schulz

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.

18.6.3 Intrusion Detection and Outbound Spam

Funding: Strato Rechenzentrum AG
Duration: since 10/2006
Project Members: Uwe Dick.

Strato AG is a major European provider of webspace and server hosting services. Intrusions are attempted on a daily basis. Usually, attackers seek to exploit insecure web sites in order to send huge amounts of spam emails via Strato's email servers. We develop an intelligent monitoring system that tracks http requests and discriminates legitimate use of a web site from attempts to exploit insecure scripts.

18.6.4 Spam: Server-Sided Identification of Spam Emails

Funding: STRATO Rechenzentrum AG
Duration: since 07/2005
Project Members: Michael Brückner, Peter Haider.

We analyze the adversarial classification problem of spam identification. Spam filtering is a game between two opponents, spam sender and spam filter, that react to each other's moves. We seek to identify a winning strategy that cannot easily be dodged by spam senders. In cooperation with Strato AG, we have developed a spam filter that now processes roughly 1 percent of all emails sent and received worldwide.

18.6.5 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

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.

18.6.6 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 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.

Learning for Information Retrieval

Both, Department 5 and Research Group 2 address personalized ranking and collaborative filtering problems. The groups collaborate on these fields.

Miscellaneous Cooperations

Members of the group have recently co-authored papers with Thomas Gärtner and Stefan Wrobel, Fraunhofer IAIS, Alexander Zien, MPI for Biological Cybernetics, Szymon Jaroszewicz, Polish National Institute of Telecommunications, Lenka Ivantysynova, Humboldt-Universität zu Berlin, David Vogel, AI Insight, Inc., and other cooperation partners. Tobias Scheffer is a member of the Interdisciplinary Center for Ubiquitous Information at Humboldt-Universität zu Berlin, and the Interdisziplinäres Zentrum für Sprachliche Bedeutung at Humboldt-Universität zu Berlin.

18.7 Publications

The following articles have been published between January 2007 and March 2007.

Journal Article

[1] S. Jaroszewicz, L. Ivantysynova, and T. Scheffer. Accurate schema matching on streams. *Intelligent Data Analysis*, 2007. In print.

Conference articles

[1] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the International Conference on Machine Learning*, 2007.

[2] 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*, Cambridge, USA, 2007, vol. 19. MIT Press.

[3] M. Brückner, P. Haider, and T. Scheffer. Highly scalable discriminative spam filtering. In *Proceedings of the 15th Text Retrieval Conference (TREC)*, 2007.

[4] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the International Conference on Machine Learning*, 2007.

[5] P. Haider, U. Brefeld, and T. Scheffer. Supervised clustering of streaming data for email batch detection. In *Proceedings of the International Conference on Machine Learning*, 2007.

19 The Discrete Optimization Group (IRG2)

19.1 Personnel

Head of Independent Research Group 2
Prof. Dr. Friedrich Eisenbrand (–December 2006)

Post-doctoral fellows and long-term guests
Christian Lennerz (–August 2005)

PhD Students
Markus Behle
Edda Happ, member of the graduates studies programme of UdS
Andreas Karrenbauer
Sören Laue
Gennady Shmonin (–March 2006)

Secretary
Ellen Fries

19.2 Integer Programming

Integer programming is today's most important industry strength method to solve difficult discrete optimization problems. Our group contributes both with solutions to theoretical problems in this area, as well as with efficient software for integer programming, especially with new tools for counting and enumerating 0/1 points in polytopes.

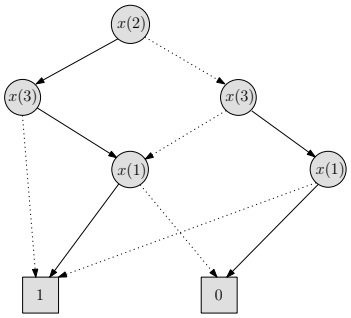
19.2.1 0/1 vertex and facet enumeration with BDDs

Investigators: Markus Behle and Friedrich Eisenbrand

In integer programming an important part in understanding and designing algorithms for a certain problem is the investigation of the polyhedral structure of the associated polytope. For many problems in this field the underlying polytope is a 0/1 polytope, i.e. all vertices are 0/1 points.

One frequently arising problem is the *0/1 vertex enumeration* problem:

> Given a set of inequalities $Ax \leq b$, $A \in \mathbb{Z}^{m \times d}$, $b \in \mathbb{Z}^m$, compute a list of all 0/1 points satisfying the system.



(a) BDD

$x(1)$	$x(2)$	$x(3)$
0	0	1
0	1	0
1	0	0
1	1	0
1	1	1

(b) Truth table

Figure 19.1: A simple BDD represented as a directed graph. Edges with parity 0 are dashed. The table shows the represented 0/1 points of the set T.

In other words, if P denotes the polyhedron $P = \left\{x \in \mathbb{R}^d \mid Ax \leq b,\ 0 \leq x \leq 1\right\}$ then one is interested in the vertices of the *integer hull* P_I of P which generate the convex hull of all integer points of P.

The other problem which we consider here is the *0/1 facet enumeration* problem:

> Given a set $S \subseteq \{0,1\}^d$ of 0/1 points, enumerate all facets of the convex hull $\mathrm{conv}(S)$.

A successful approach to difficult optimization problems with integer programming often requires some understanding of the facets of the integer hull of the solution space. A software package which computes the inequality representation $A'x \leq b'$ of the integer hull P_I of P, given an inequality representation $Ax \leq b$ of P can here become very useful. Such an inequality representation is currently computed in a two-step approach. In a first step, one solves the 0/1 vertex enumeration problem, and then in a second step the 0/1 facet enumeration problem for the previously generated 0/1 points is solved.

In [2] we develop tools to solve the 0/1 vertex and facet enumeration problems which are based on binary decision diagrams (BDDs), see figure 19.1. Our freely available tool `azove` [1] which solves the vertex enumeration problem outperforms the currently best codes for this task by several orders of magnitude. Second we report on a gift-wrapping approach to solve the facet enumeration problem. Here we use BDDs to rotate a facet-defining inequality along a ridge to find a new facet. Ridges are computed with existing codes. We can recommend our approach for polytopes whose facets contain few vertices.

References

- [1] M. Behle. *Another Zero One Vertex Enumeration tool Homepage*, 2006. http://www.mpi-inf.mpg.de/∼behle/azove.html.
- [2] M. Behle and F. Eisenbrand. 0/1 vertex and facet enumeration with BDDs. In *Workshop on Algorithm Engineering and Experiments (ALENEX'07)*, New Orleans, U.S.A., 2007. to appear.

19.3 Combinatorial Optimization

Combinatorial optimization is a core-discipline in algorithmic discrete mathematics and complexity. We are concerned with the design and analysis of efficient algorithms for tractable problems and with the design of approximation algorithms for NP-hard optimization problems.

19.3.1 Algorithms for longer OLED Lifetime

Investigators: Friedrich Eisenbrand and Andreas Karrenbauer

We consider an optimization problem arising in the design of controllers for OLED displays in a joint project with the electrical engineering department of Saarland University. Our objective is to minimize the amplitude of the electrical current flowing through the diodes which has a direct impact on the lifetime of such a display. The optimization problem consist of finding a decomposition of an image into subframes with special structural properties that allow the display driver to lower the stress on the diodes. Modeling the problem in mathematical terms yields a class of network flow problems where we group the arcs and pay in each group only for the arc with the highest payload. We developed (fully) combinatorial approximation heuristics suitable for being implemented in the hardware of a control device that drives such an OLED display [1].

Organic Light Emitting Diodes (OLEDs) have received growing interest recently as more and more commercial products are equipped with such displays. Though they have many advantages over current technology like LCD, only small size OLED displays have entered the marked yet. One reason for this is the limited lifetime of those displays. While a lot of research is conducted on the material science side, we invented the so-called *Multiline Addressing Scheme* for passive matrix OLED displays [3] that tackles this problem from a driver point of view. It is based on the fact that equal rows can be displayed simultaneously with a lower electrical current than in a serial manner. Therefore, we have to compute a decomposition of an image (see Fig. 19.3.1 for an example) that permits the driver to exploit that property [2].

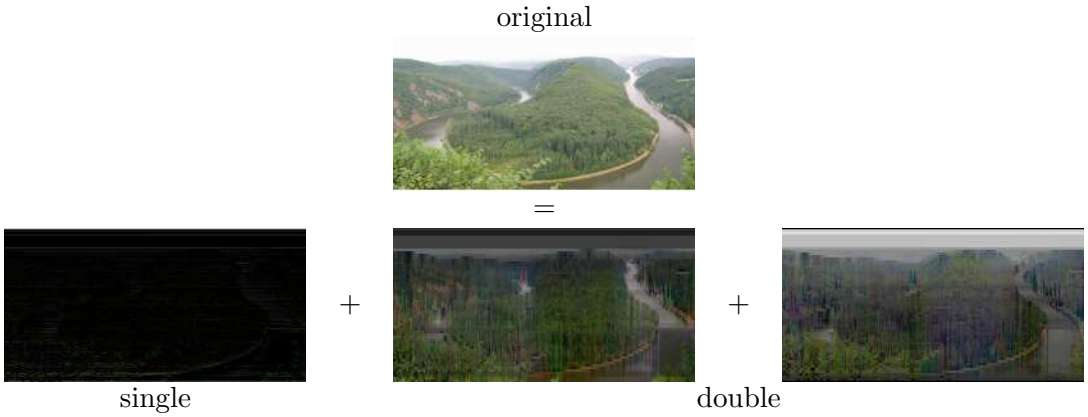


Figure 19.2: Decomposition of an image such that every two rows of the double parts have the same content.

To benefit from such a decomposition in practice, an algorithm for this optimization problem has to be implemented on a chip which is attached to the display. Therefore, the following design criteria lie in the focus when engineering such an algorithm.

- The algorithm has to react in realtime.
- The algorithm must have low hardware complexity allowing small production costs.
- Consequently it has to rely only on a small amount of memory and it should be *fully combinatorial*, i.e. only additions, subtractions, and comparisons are used.

We accomplished all these goals and came up with an algorithm that allows a very economic hardware implementation that takes only roughly 10000 gates for a display of 100×100 pixels. Moreover, it achieves near optimal solutions on real world images. Thereby, the lifetime improves significantly or equivalently 3–10 times larger displays with the same expected lifetime become possible.

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] C. Xu, A. Karrenbauer, K. M. Soh, and J. Wahl. A new addressing scheme for PM OLED display. *Society for Information Display (SID) Symposium Digest*, to appear, 2007.
- [3] C. Xu, J. Wahl, F. Eisenbrand, A. Karrenbauer, K. M. Soh, and C. Hitzelberger. Verfahren zur Ansteuerung von Matrixanzeigen. *Patent 10 2005 063 159, Germany, pending*, 2005.

19.3.2 Virtual Private Network Design

Investigators: Friedrich Eisenbrand and Edda Happ

We studied the *virtual private network design* (VPND) problem which arises e.g. if one wants to connect different company sites over the internet using a virtual private network. The problem is specified as follows: Given an undirected, weighted graph and a set of terminal vertices, specify paths between all terminal pairs and reserve capacities of minimum cost on the edges. The solution has to assure that all matchings of terminals can be routed along the corresponding paths without exceeding the reserved capacities. Since every terminal can communicate with any other terminal the problem is also called the symmetric VPND problem, whereas in the asymmetric version the terminals are divided into two sets (senders and receivers) and only terminals of different groups can communicate. The asymmetric VPND models networks with client/server applications.

In [1] we considered a generalization of both VPND versions where we have k sets of terminals and gave a 4.74 approximation algorithm for this problem.

References

- [1] F. Eisenbrand and E. Happ. Provisioning a virtual private network under the presence of non-communicating groups. In T. Calamoneri, I. Finocchi, and G. F. Italiano, eds., *Algorithms and complexity, 6th Italian Conference, CIAC 2006*, Rome, Italy, 2006, *LNCS 3998*, pp. 105–114. Springer.

19.4 Dissertations

19.4.1 Dissertations in progress

Markus Behle, *Binary Decision Diagrams and Integer Programming*
Andreas Karrenbauer, *Multiline Addressing by Network Flow*

19.5 Grants and Cooperations

19.5.1 AVACS

Involved Persons: Markus Behle, Friedrich Eisenbrand, Gennady Shmonin

Our research group is part of the Transregional Collaborative Research Center 14 AVACS (Automatic Verification and Analysis of Complex Systems), which is a SFB funded by the German Research Foundation (DFG). This project addresses the mathematical analysis of models of complex safety critical computerized systems, such as aircrafts, trains, cars, or other artifacts. The overall aim is to raise the state of the art in automatic verification and analysis techniques to a level allowing a comprehensive verification of such systems.

We provide the knowledge for linear and integer programming and combine our methods with techniques from verification and satisfiability. Apart from the MPII, the Universities of Freiburg, Oldenburg and Saarbrücken take part in this project. For more details see `www.avacs.org`.

19.6 Publications

Book chapters

[1] F. Eisenbrand and K. Aardal. Integer programming, lattices and results in fixed dimension. In K. Aardal, G. Nemhauser, and R. Weismantel, eds., *Discrete Optimization*, *Handbooks in Operations Research and Management Science*, vol. 12, ch. Chapter 4, pp. 171–244. Elsevier, Amsterdam, The Netherlands, 2005.

Journal articles

[1] F. Eisenbrand and S. Laue. A linear algorithm for integer programming in the plane. *Mathematical Programming*, 102(2):249–259, 2005.

[2] F. Eisenbrand and G. Shmonin. Carathéodory bounds for integer cones. *Operations Research Letters*, 34(5):564–568, 2006.

Conference articles

[1] B. Becker, M. Behle, F. Eisenbrand, and R. Wimmer. BDDs in a branch and cut framework. In S. Nikoletseas, ed., *4th International Workshop on Experimental and Efficient Algorithms (WEA'05)*, Santorini, Greece, May 2005, *LNCS 3503*, pp. 452–463. Springer.

[2] M. Behle. On threshold BDDs and the optimal variable ordering problem. In D.-Z. Du, ed., *First International Conference on Combinatorial Optimization and Applications (COCOA'07)*, Xi'an, China, August 2007, *LNCS 4616*. Springer, to appear.

[3] M. Behle and F. Eisenbrand. 0/1 vertex and facet enumeration with BDDs. In *Workshop on Algorithm Engineering and Experiments (ALENEX'07)*, New Orleans, U.S.A., 2007. to appear.

[4] 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., *6th Workshop on Experimental Algorithms (WEA'07)*, Rome, Italy, June 2007, *LNCS 4525*, pp. 379–392. Springer.

[5] F. Eisenbrand, S. Funke, A. Karrenbauer, and D. Matijevic. Energy-aware stage illumination. In *Proceedings of the Twenty-First Annual Symposium on Computational Geometry (SCG-05)*, Pisa, Italy, 2005, pp. 336–345. ACM.

[6] F. Eisenbrand, S. Funke, A. Karrenbauer, E. Schoemer, and J. Reichel. Packing a trunk - now with a twist! In S. N. Spencer, ed., *SPM 2005, ACM Symposium on Solid and Physical Modeling*, Cambridge, USA, 2005, pp. 197–206. ACM.

[7] F. Eisenbrand and F. Grandoni. An improved approximation algorithm for virtual private network design. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA 05)*, Vancouver, Canada, 2005, vol. 16, pp. 928–932. SIAM.

[8] F. Eisenbrand, F. Grandoni, G. Oriolo, and M. Skutella. New approaches for virtual private network designs. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, eds., *Automata, languages and programming, 32nd International Colloquim, ICALP 2005*, Lisboa, Portugal, 2005, *LNCS 3580*, pp. 1151–1162. Springer.

[9] F. Eisenbrand and E. Happ. Provisioning a virtual private network under the presence of non-communicating groups. In T. Calamoneri, I. Finocchi, and G. F. Italiano, eds., *Algorithms and complexity, 6th Italian Conference, CIAC 2006*, Rome, Italy, 2006, *LNCS 3998*, pp. 105–114. Springer.

[10] 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.

[11] F. Eisenbrand, A. Karrenbauer, and C. Xu. Algorithms for longer oled lifetime. In C. Demetrescu, ed., *WEA 2007*, Rome, Italy, June 2007, *LNCS 4525*, pp. 338–351. Springer.

[12] F. Eisenbrand, G. Oriolo, G. Stauffer, and P. Ventura. Circular ones matrices and the stable set polytope of quasi-line graphs. In M. Jünger and V. Kaibel, eds., *Integer programming and combinatorial optimization, 11th International IPCO Conference (IPCO XI)*, Berlin, Germany, 2005, *LNCS 3509*, pp. 291–305. Springer.

Theses

[1] J. Rieskamp. Automation and optimization of a trunk packing process. Masters thesis, Universität des Saarlandes, September 2005. Diplomurkunde 26.04.2006.

20 The Graphics–Optics–Vision Group (IRG3)

20.1 Personnel

Director
Prof. Dr. Ing. Marcus Magnor (–March 2007)

Researchers
Lukas Ahrenberg
Ivo Ihrke
Andrei Lintu
Christian Linz (–April 2006)
Volker Scholz
Timo Stich (–February 2006)

Secretary
Ellen Fries

20.2 Visitors

In the time period from March 2005 to January 2006, the following researchers visited our group:

Kiriakos Kutulakos	23.10.05–26.10.05	University of Toronto

20.3 Group Organization

With the appointment of Marcus Magnor as full professor to TU Braunschweig in January 2006, the Independent Research Group "Graphics–Optics–Vision" was merged with the Computer Graphics Department under the supervision of Hans-Peter Seidel. Two PhD students of the group came along to Braunschweig, while the four more senior members decided to stay in Saarbrücken. All four remaining PhD students are now officially supervised by Hans-Peter Seidel, with Marcus Magnor as co-supervisor. By courtesy of Saarland University, Marcus Magnor was allowed to remain principal supervisor of all theses that were finished in 2006. Among these is the PhD thesis of the first member of the group who received his doctoral degree in July 2006, with distinction. In addition, several Bachelor and Master students of the group were able to finish their theses during 2006. The newly established

group at TU Braunschweig keeps cordial ties to MPII and pursues several research projects in close cooperation with the Computer Graphics Department D4.

20.4 Research Projects

20.4.1 Computer Generated Holograms and Computational Holography

Digital holography has grown strong recently as better and better CCD chip technology has become available. At the same time, the introduction of high resolution Spatial Light Modulators have made holographic display technology a possibility. While digital holography still can not compare to traditional holography in resolution, several unique features such as phase-shift techniques and faster recording procedures have made it widely used in many areas, including image processing. Computers are also used to synthesize wave front patterns for holographic displays.

Computer generated holography from 3D models

Investigator: Lukas Ahrenberg

One possible way of realizing a holographic display is through the use of Spatial Light Modulators (SLMs). The performance and resolution of SLMs are increasing rapidly; consequently, holographic displays of increasing spatial bandwidth product can be built. To drive these, Computer Generated Holograms (CGHs), rendered from point models are commonly used. However, generating a hologram from a set of point samples is a computationally intensive task. Within this project we have proposed a method that takes advantage of parallel graphic processing units (GPUs) to perform the computation. [1]. Although development of special purpose hardware for CGH rendering has been reported, this type of equipment is expensive and must be custom built. In contrast, graphic boards are readily available today, and are specially constructed to accelerate numerical operations frequently used in computer graphics. Thus, a parallel GPU system is an attractive alternative to both expensive custom-built hardware and too much slow CPU-based approaches.

The approach used in [1] propagate the light from each 3D object to each image point using a special formulation of the Rayleigh Sommerfeld diffraction formula. Previous GPU-based methods have either used pre-computed Fresnel zone plates, making one pass per object point, or implemented the light transport formula in a fragment shader. Both these approaches have disadvantages. The zone-plates are approximations, and will introduce errors if the depth value of the object point differs from the zone-plate depth. Implementing the light transport formula simply in a fragment shader requires looping which is both restrictive and inefficient. Our approach estimates the number of points that will be rendered and generates a fragment-shader on the fly by unrolling the loops manually, performing just so many passes that are needed. The implementation show good results and efficiency.

However effective, this approach shares a specific weakness in common with most CGH rendering algorithms: while the 3D model is described by parametric surfaces, or by polygon patches it has to be sampled to point sources before light transport. The resulting wave field is then sampled again at the image plane. This is a drawback compared to traditional computer graphics where the sole sampling is in the image plane. We are currently working

on the next stage of this project considering an analytic approach that should reduce the problem from per-point rendering to a per-surface.

References

- [1] L. Ahrenberg, P. Benzie, M. Magnor, and J. Watson. Computer generated holography using parallel commodity graphics hardware. *Optics Express*, 14(17):7636–7641, August 2006.

Connecting Holograms and Light Fields

Investigator: Lukas Ahrenberg

In contrast to a photograph, a hologram represents a sampling of the complete light emitted from the scene. Therefore, digital holography has the potential to open up new possibilities in image analysis and computer vision. Hologram analysis, much like multi-camera based analysis is not confirmed to a single viewpoint, but contains 3D information. A recorded hologram can be regarded as sampling the light passing through an aperture. The representation however, is not straightforward from a computational perspective, as it stores the complex wave front of light and thus no explicit depth information. However, if the information encoded in holograms can be effectively accessed and utilized it would prove a good foundation for extended scene analysis, and might open up to novel vision algorithms.

From a computer graphics perspective the holographic concept is similar to the light field. Both capture the full light from a scene as it passes through a plane in space. We have therefore started to investigate how these two representations are related.

Although their conceptual similarities, the hologram and light field representations differ on a fundamental level. The hologram samples the complex valued wave front, while the light field is a ray-based representation sampling a pencil of rays passing through a point in space. This fundamental difference may mean that an explicit 3D reconstruction in order to determine depth values may be required to convert fully between the representations.

While a complete one to one mapping remains to be explored, we have started to analyze holograms from a light field perspective using time-frequency methods. In short, the wave front encoded in a hologram has perfect spatial location, but nothing is known on the light directions. The Fourier transform of the wave front, its angular spectrum, can however be physically interpreted as plane waves. These have perfect angular localization, but by Fourier transforming the spatial information is lost. Time-frequency analysis employs methods where a local spectra is reconstructed, containing both spatial and angular localization. This process is associated with an uncertainty function, as perfect localization in both space and frequency is not possible to attain.

We have performed initial testing using the short-term Fourier transform with good results, and are planning to investigate more advanced time-frequency methods such as the Wiegner transform in the future.

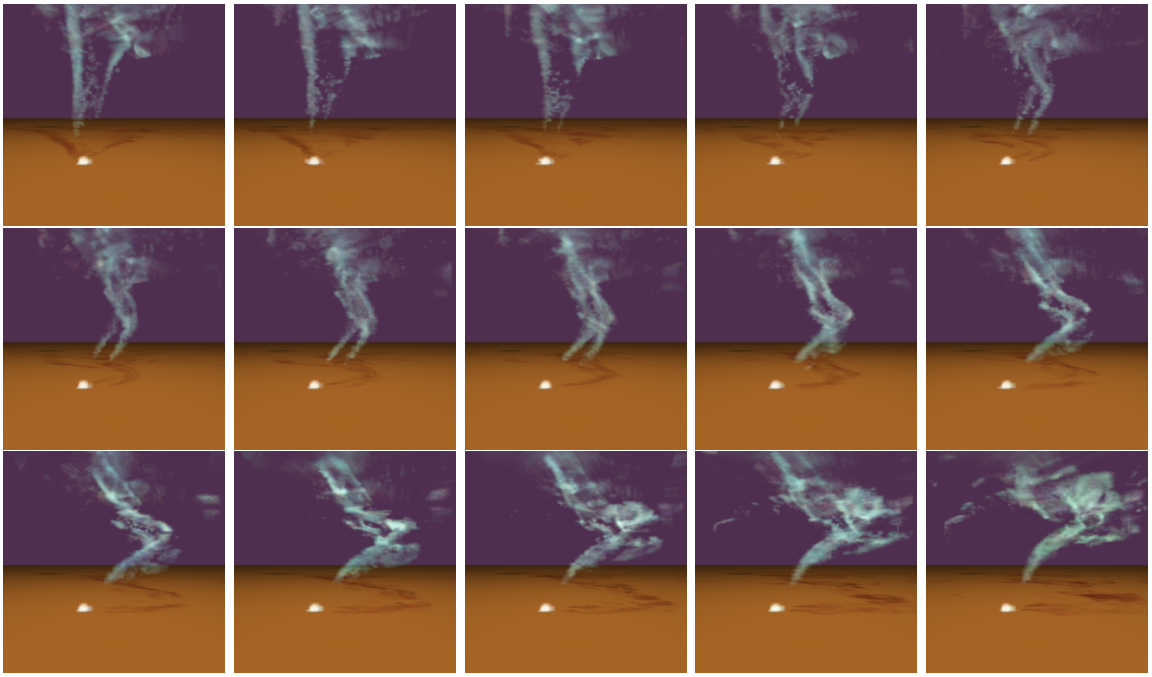


Figure 20.1: Reconstructions of 15 consecutive frames of a smoke sequence.

20.4.2 Adaptive Grid Tomography

Investigator: Ivo Ihrke

In this project we investigated extensions to earlier work on tomographic reconstruction of flames [1]. We applied the developed techniques to the reconstruction of thin smoke columns. The extensions to the original algorithm are mostly technical in nature. We introduced an octree-based adaptive grid technique to be able to resolve fine details of the whirling smoke columns. Previously we were restricted by the main memory of personal computers when using a uniform discretization. An example of the results is shown in Fig. 20.1. The use of an adaptive representation for the volumetric models however resulted in performance and regularization problems for the adaptive algorithm. We investigated the causes for this and developed new solution and regularization methods to deal with these problems. On the solution side we investigated the use of preconditioners for large, sparse linear systems. We developed a method to transfer error measures defined in the image plane to the three-dimensional reconstruction grid. On the regularization side we investigated different explicit regularization methods. We found problems with the standard approaches in the literature and developed an adaptive version of the Tikhonov regularization method. The details of the discussed improvements can be found in [2, 3].

References

- [1] I. Ihrke and M. Magnor. Image-Based Tomographic Reconstruction of Flames. In *ACM Siggraph / Eurographics Symposium Proceedings, Symposium on Computer Animation*, 2004, pp. 367–375.

- [2] I. Ihrke and M. Magnor. Adaptive grid optical tomography. In D. Fellner, ed., *Vision, Video, and Graphics 2005*, Edinburgh, UK, 2005, pp. 141–147. Eurographics.
- [3] I. Ihrke and M. Magnor. Adaptive grid optical tomography. *Graphical Models*, 68(5-6):484–495, 2006.

20.4.3 Augmented Astronomical Telescope

Investigators: Andrei Lintu and Marcus Magnor

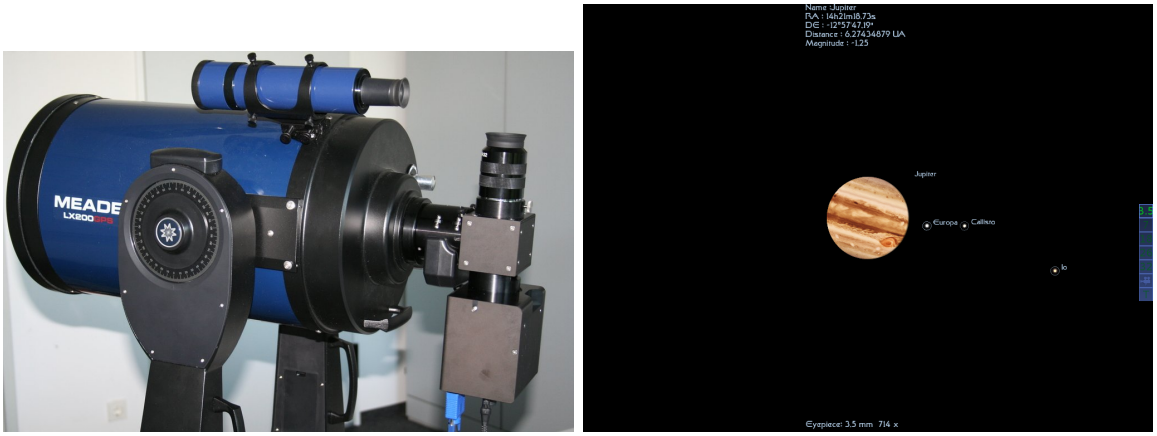


Figure 20.2: Left: The projection unit mounted on the telescope. Right: A snapshot of the view through the telescope's eyepiece, while observing Jupiter and it's Moons. Additional information about the observed object as well as a label to ease the identification of the planets Moons is highlighted

Anyone who gazed through the eyepiece of an astronomical telescope knows that except for the Moon and the planets, extra-solar astronomical objects are hard to observe. This is mainly due to their low surface brightness, but also depends on the seeing, sky brightness and telescope aperture. We propose a system [1, 2] which projects images of astronomical objects (with focus on nebulae and galaxies), animations or additional information directly into the eyepiece view of an astronomical telescope. As the telescope orientation is tracked continuously, the projected image is adapted in real-time to the object which is currently visible through the eyepiece. This way, visitors to public observatories have the possibility to experience the richness of deep sky objects while directly gazing at them through a telescope.

References

- [1] A. Lintu and M. Magnor. Augmented astronomical telescope. In T. Kuhlen, L. Kobbelt, and S. Müller, eds., *Virtuelle und erweiterte Realität, 2. Workshop der GI-Fachgruppe VR/AR*, Aachen, Germany, 2005, pp. 203–213. Shaker.
- [2] A. Lintu and M. Magnor. An augmented reality system for astrnomical observations. In B. Froelich, D. Bowman, and H. Iwata, eds., *IEEE Virtual Reality 2006*, Alexandria, Virginia, USA, 2006, pp. 119–126. IEEE.

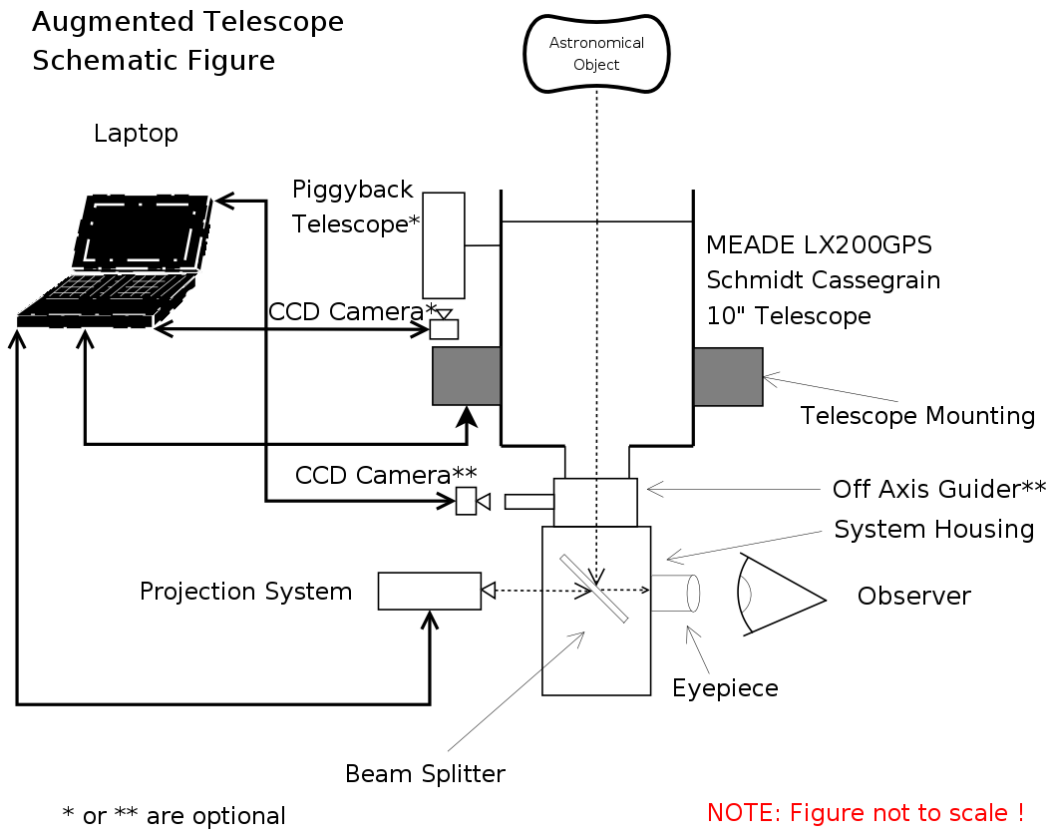


Figure 20.3: Schematic description of the entire system. It's main hardware components are the astronomical telescope, a custom projection unit and a personal computer.

20.4.4 Effective Multi-resolution Rendering and Texture Compression for Captured Volumetric Trees

Investigators: Christian Linz and Marcus Magnor

Modeling and rendering trees has been a goal of computer graphics research since the early days of the field.

While most of the effort has been in solutions to generate entirely synthetic trees (e.g., [1]), an alternative is the approach to capture and render real trees [3]. For both synthetic and captured trees, however, polygonal representations (mainly of the leaves) result in objects which are very complex and thus expensive to render. In addition, generating geometric levels-of-detail (LOD) for disconnected triangle meshes, such as the leaves of a tree, is an unsolved problem; the few solutions proposed to date require mixing various different representations. However, trees

are a good candidate for volumetric representations [3]; one big advantage of such an approach are appropriate multi-resolution LOD structures resulting naturally from the hierarchical data structure representing the volume.

Although Reche et al. [3] did use a volumetric representation, no multi-resolution solution was presented, and the texture memory requirements were prohibitively high. Despite the realistic renderings provided by the approach, the method remains unusable for all practical purposes (60,000-140,000 polygons and 60-140MB of texture memory per tree).

Our work [2] is motivated by those shortcomings. Our method uses an octree with appropriate textures at intermediate hierarchy levels and applies an effective pruning strategy. For texture compression, we adapt a vector quantization approach in a perceptually accurate color space, and modify the codebook generation of the Generalized Lloyd Algorithm to further improve texture quality. In combination with several hardware acceleration techniques, our approach achieves a reduction in texture memory requirements by two orders of magnitude compared to the method of Reche et al. [3]. In addition, it is now possible to render tens or even hundreds of captured trees at interactive rates.

References

[1] O. Deussen, P. Hanrahan, B. Lintermann, R. Měch, M. Pharr, and P. Prusinkiewicz. Realistic modeling and rendering of plant ecosystems. *Proc. SIGGRAPH'98*, pp. 275–286, 1998.

- [2] C. Linz, G. Drettakis, M. Magnor, and A. Reche-Martinez. Effective multi-resolution rendering and texture compression for captured volumetric trees. In E. Galin and N. Chiba, eds., *Proceedings of the Eurographics Workshop on Natural Phenomena*, Vienna, Austria, September 2006, pp. 83–90. Eurographics.

[3] A. Reche-Martinez, I. Martín, and G. Drettakis. Volumetric reconstruction and interactive rendering of trees from photographs. *ACM Trans. Graph.*, 23(3):720–727, 2004.

20.4.5 Reflection Nebula Visualization

Investigators: Marcus Magnor, Kristian Hildebrand, and Andrei Linţu

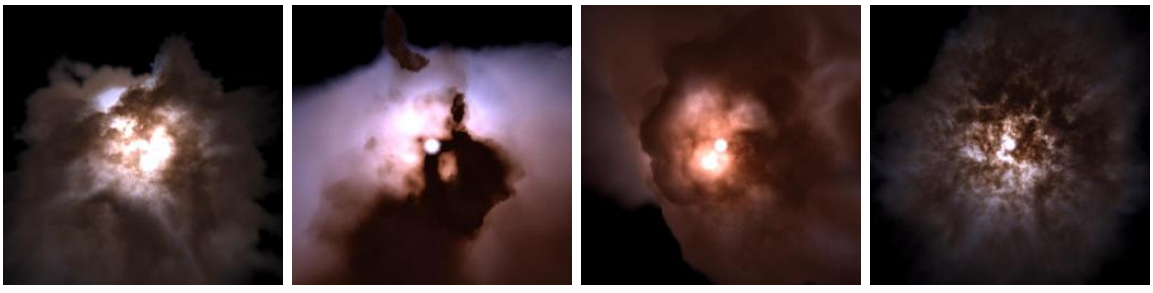


Figure 20.4: Several views of synthetic datasets using our physically correct volume renderer

Stars form in dense clouds of interstellar gas and dust. The residual dust surrounding a young star scatters and diffuses its light, making the star's "cocoon" of dust observable from Earth. The resulting structures, called reflection nebulae , are commonly very colorful in appearance due to wavelength-dependent effects in the scattering and extinction of light.

The intricate interplay of scattering and extinction cause the color hues, brightness distributions, and the apparent shapes of such nebulae to vary greatly with viewpoint. We have developed an interactive visualization tool [1] for rendering physically correct representations of arbitrary dust distributions surrounding a central illuminating star. Our algorithm can be used to create virtual fly-throughs of a reflection nebula for either interactive desktop visualizations or scientifically accurate animations for planetarium shows. Besides being able to take the measured data from a given reflection nebula as input and producing a plausible simulation of the nebula's appearance from arbitrary viewpoints, the system also supports the investigation of the visual effects of changing a nebula's physical parameters. Our tool permits the exploration of alternate plausible models for reflection nebulae, and exploit visualization to gain a deeper and more intuitive understanding of the complex interaction of light and dust in real astrophysical settings.

References

- [1] M. Magnor, K. Hildebrand, A. Lintu, and A. J. Hanson. Reflection nebula visualization. In C. Silva, E. Gröller, and H. Rushmeier, eds., *IEEE Visualization 2005, VIS 2005*, Minneapolis, USA, 2005, pp. 255–262. IEEE.

20.4.6 Garment Motion Capture Using Color-Coded Patterns

Investigators: Volker Scholz, Timo Stich, and Marcus Magnor

In this project [1, 2] we present an image-based algorithm for surface reconstruction of moving garment from multiple calibrated video cameras. Using a color-coded cloth texture, we reliably match circular features between different camera views. As surface model we use an a priori known triangle mesh. By identifying the mesh vertices with texture elements we obtain a consistent parameterization of the surface over time without further processing. Missing data points resulting from self-shadowing are plausibly interpolated by minimizing a thin-plate functional. The deforming geometry can be used for different graphics applications, e.g. for realistic retexturing. We show results for real garments demonstrating the accuracy of the recovered flexible shape.

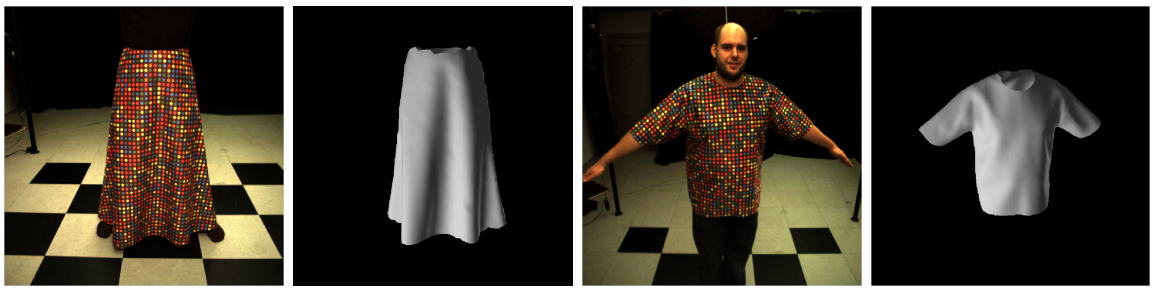


Figure 20.5: Input frames and reconstruction results for a skirt and a T-shirt.

References

- [1] V. Scholz, T. Stich, M. Keckeisen, M. Wacker, and M. Magnor. Garment motion capture using color-coded patterns. *Computer Graphics forum*, 24(3):439–448, September 2005. Conference Issue: 26th annual Conference Eurographics 2005, Dublin, Ireland, August 29th - September 2nd, 2005.

- [2] V. Scholz, T. Stich, M. Magnor, M. Keckeisen, and M. Wacker. Garment motion capture using color-coded patterns. In J. Buhler, ed., *ACM SIGGRAPH Sketches and Applications*, Los Angeles, USA, August 2005, p. 1. ACM SIGGRAPH.

20.4.7 Texture Replacement of Garments in Monocular Video Sequences

Investigators: Volker Scholz and Marcus Magnor

This project addresses a specific problem of movie post-production. We propose a system to enable texture replacement with correct texture deformation and lighting. We present a video processing algorithm for texture replacement of moving garments in monocular video recordings in [1]. We use a color-coded pattern which encodes texture coordinates within a local neighborhood in order to determine the geometric deformation of the texture. A time-coherent texture interpolation is obtained by the use of 3D radial basis functions. Shading maps are determined with a surface reconstruction technique and applied to new textures which replace the color pattern in the video sequence. Our method enables exchanging fabric pattern designs of garments worn by actors as a video post-processing step.

Figure 20.6: Input frame (left) and three texture replacement results.

References

- [1] V. Scholz and M. Magnor. Texture replacement of garments in monocular video sequences. In D. W. Fellner, S. N. Spencer, Y. Chrysanthou, D. Cohen-Or, T. Akenine-Möller, and W. Heidrich, eds., *Rendering Techniques 2006, Eurographics Symposium on Rendering*, Nicosia, Cyprus, June 2006, pp. 305–312. Eurographics.

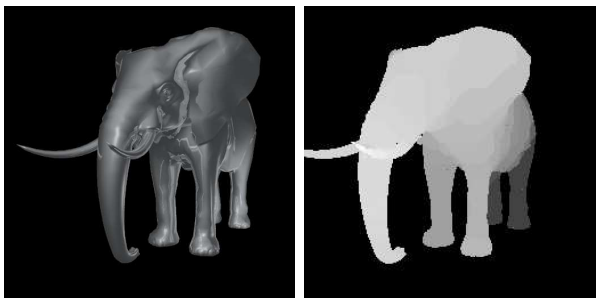


Figure 20.7: Left: One of the input images of a highly specular object. Right: Reconstructed depth image. Shiny and occluded parts are reconstructed.

20.4.8 Global Depth from Epipolar Volumes – A General Framework for Reconstructing Non-Lambertian Surfaces

Investigators: Timo Stich, Art Tevs, and Marcus Magnor

Solving the passive multi-view 3D reconstruction problem has and still is one of the most worked on problems in the computer vision community. Motivated by the human ability to easily perceive a 3D world with only two "cameras", many different approaches have been developed. However, most approaches rely on strong assumptions on the BRDF of the scene objects e.g. to be lambertian which is not true in general. For example shiny materials like plastic or metals violating this assumption cause artifacts and errors in the reconstruction results. We developed a novel view of the correspondence finding problem by using a subset of the Plenoptic Function, the Epipolar Volume [1]. Using Epipolar Image Analysis in the context of the correspondence finding problem in depth reconstruction has several advantages. One is the elegant incorporation of prior knowledge about the scene or the surface reflection properties into the reconstruction process. The proposed framework in conjunction with graph cut optimization [2] is able to reconstruct also highly specular surfaces. The use of prior knowledge in general opens up new ways to reconstruct complicated surfaces and scenes impossible with previous methods. Another advantage is improved occlusion handling, which also allows pixels that are partly occluded to contribute to the reconstruction results. The proposed shifting of some of the computation to graphics hardware (GPU) results in a significant speed improvement compared to pure CPU-based implementations.

References

[1] M. M. T Stich, A. Tevs. Global depth from epipolar volumes - a general framework for reconstructing non-lambertian surfaces. In *Third International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2006, pp. 1–8.

[2] O. V. Y. Boykov and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. In *IEEE Transactions on pattern analysis and machine intelligence*, 2001, pp. 1222–1239.

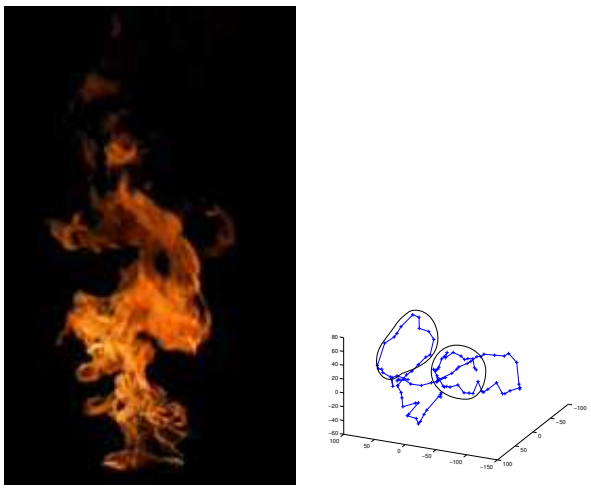


Figure 20.8: Left: Image from a video sequence processed with our method. Right: Proposed visualization of the sequence which reveals periodic patterns.

20.4.9 Keyframe Animation from Video

Investigators: Timo Stich and Marcus Magnor

Image based approaches are advantageous to use if synthetic scene models are time-consuming to create or do not achieve satisfying realism. Examples are dynamic natural phenomena like fire, smoke and water. However, the drawback of image sequences is the lack of manipulative freedom. We propose a method for analyzing and synthesizing video sequences, specifically suited for image sequences of natural phenomena [2]. We combine a low-dimensional representation of arbitrary image sequences and an image morphing technique to create realistic in-between images.

In a first step, we address the problem of selecting snippets having a periodic character. The visualization based on the Isomap algorithm [3] allows users to easily select these parts. In a second step, these subsequences are then in turn used to synthesize new sequences with the desired properties in real-time. To smooth transition artifacts between the reordered subsequences a Monge-Kantorovich based image morphing method [1] is applied to interpolate in-between images. Users can easily create looping versions of the input sequence, script the appearance over time or adjust the temporal resolution.

References

[1] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal mass transport for registration and warping. *International Journal on Computer Vision*, 60(3):225–240, 2004.

[2] M. M. T. Stich. Keyframe animation from video. In *International Conference on Image Processing (ICIP)*, 2006, pp. 2713–2716.

[3] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

20.5 Academic Activities

Membership in Program Committees

Marcus Magnor:

- *Eurographics (EG) 2006*
- *Eurographics Symposium on Rendering (EGSR) 2006*
- *IEEE Workshop on Three-Dimensional Cinematography (3DCINE) 2006*
- *International Conference on Image Processing (ICIP) 2006*
- *International Conference in Central Europe on Computer Graphics, Visualization, and Computer Vision (WSCG) 2005, 2006*
- *International Conference on Computer Graphics Theory and Applications 2006*
- *International Symposium on 3D Data Processing, Visualization and Transmission, (3DPVT) 2006*
- *International Workshop on Volume Graphics, (VG) 2006*
- *Vision, Modeling, and Visualization (VMV) 2005, 2006*

Membership in Organizing Committees

Marcus Magnor:

- *European GAME-ON Conference 2006*

20.6 Teaching Activities

Winter Semester 2005/2006

Seminars:
Data Processing Tips and Tricks (M. Magnor)

Summer Semester 2005

Courses:
3D Image Analysis and Synthesis (V. Blanz, M. Gösele, M. Magnor)

Seminars:
Computer Game Development (M. Magnor)

Master Theses

Kristian Hildebrand: Rendering and Reconstruction of Astronomical Objects, 2005.
Christian Linz: PDE-based Surface Reconstruction from Multiple Views using a Surfel Model, 2005.

Fabian Recktenwald: Sampling and interpolation of the plenoptic function from sparse data, 2006.

Bachelor Theses

Sascha El-Abed: A computer graphics-based analysis-by-synthesis approach for the joint recovery of the gas and dust distributions in bipolar planetary nebulae from visible and radio wavelength observations, 2006.

Martin Schaef: Rekonstruktion von bewegtem Stoff mittels direkter Suchverfahren, 2005.

20.7 Dissertations, Habilitations, Offers, Awards

20.7.1 Dissertations

Bastian Goldluecke: Multi-Camera Reconstruction and Rendering for Free-Viewpoint Video, 2006.

20.7.2 Habilitations

Marcus Magnor: Visuelle Ästhetik vs. Visual Computing, 2005.

20.7.3 Offers for Faculty Positions

Marcus Magnor: TU Braunschweig, 2005.

20.8 Grants and Cooperations

3DTV – Integrated Three-Dimensional Television – Capture, Transmission, and Display

Developing complete systems for acquisition, compression, and visualization of real-world scenes for 3D display

09/2004-08/2008

European Union, 6. IST Frame Programme

Pays for: Lukas Ahrenberg

Other MPII participants: Marcus Magnor, Volker Scholz, Gernot Ziegler, Edilson de Aguiar, Naveed Ahmed

Partners:

- Bilkent University, Turkey
- Heinrich-Hertz-Institut, Berlin
- Tü¡bingen University
- TU Illmenau
- ...

20.9 Publications

Books

[1] M. Magnor. *Video-based Rendering.* A K Peters, Wellesley, USA, 2005.

Journal articles and book chapters

[1] L. Ahrenberg, P. Benzie, M. Magnor, and J. Watson. Computer generated holography using parallel commodity graphics hardware. *Optics Express*, 14(17):7636–7641, August 2006.

[2] B. Goldluecke, I. Ihrke, C. Linz, and M. Magnor. Weighted minimal surface reconstruction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2007. to appear.

[3] J. Haber, M. Magnor, and H.-P. Seidel. Physically based simulation of twilight phenomena. *Transactions on Graphics*, 24(4):1353–1373, October 2005.

[4] I. Ihrke and M. Magnor. Adaptive grid optical tomography. *Graphical Models*, 68(5-6):484–495, 2006.

[5] M. Magnor, D. Kindlmann, C. Hansen, and N. Duric. Reconstruction and visualization of planetary nebulae. *IEEE Transactions on Visualization and Computer Graphics*, 11(5):485–496, September 2005.

Conference articles

[1] N. Ahmed, E. de Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel. Automatic generation of personalized human avatars from multi-view video. In *VRST '05: Proceedings of the ACM symposium on Virtual reality software and technology*, Monterey, USA, December 2005, pp. 257–260. ACM.

[2] L. Ahrenberg, I. Ihrke, and M. Magnor. Volumetric reconstruction, compression and rendering of natural phenomena from multi-video data. In E. Gröller, I. Fujishiro, K. Müller, and T. Ertl, eds., *Volume graphics 2005, Eurographics/IEEE VGTC workshop proceedings*, Stony Brook, New York, USA, 2005, pp. 83–90. Eurographics.

[3] L. Ahrenberg and M. Magnor. Light field rendering using matrix optics. *Journal of WSCG*, 14(1-3):177–184, 2006. 14th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic, January 2006.

[4] B. Goldluecke and M. Magnor. Space-time continuous geometry meshes from multi-view video sequences. In *2005 International Conference on Image Processing (ICIP)*, Genova, Italy, September 2005, vol. 1, pp. 625–628. IEEE.

[5] I. Ihrke, B. Goldluecke, and M. Magnor. Reconstructing the geometry of flowing water. In *VISION '05, proceedings of the 2005 International Conference on Computer Vision (ICCV'05)*, Beijing, China, 2005, vol. II, pp. 1055–1060. IEEE.

[6] I. Ihrke and M. Magnor. Adaptive grid optical tomography. In D. Fellner, ed., *Vision, Video, and Graphics 2005*, Edinburgh, UK, 2005, pp. 141–147. Eurographics.

[7] A. Lintu and M. Magnor. Augmented astronomical telescope. In T. Kuhlen, L. Kobbelt, and S. Müller, eds., *Virtuelle und erweiterte Realität, 2. Workshop der GI-Fachgruppe VR/AR*, Aachen, Germany, 2005, pp. 203–213. Shaker.

[8] A. Lintu and M. Magnor. An augmented reality system for astrnomical observations. In B. Froelich, D. Bowman, and H. Iwata, eds., *IEEE Virtual Reality 2006*, Alexandria, Virginia, USA, 2006, pp. 119–126. IEEE.

[9] C. Linz, G. Drettakis, M. Magnor, and A. Reche-Martinez. Effective multi-resolution rendering and texture compression for captured volumetric trees. In E. Galin and N. Chiba, eds., *Proceedings of the Eurographics Workshop on Natural Phenomena*, Vienna, Austria, September 2006, pp. 83–90. Eurographics.

[10] C. Linz, B. Goldluecke, and M. Magnor. A point-based approach to pde-based surface reconstruction. In K. Franke, K. R. Müller, B. Nickolay, and R. Schäfer, eds., *Pattern Recognition, 28th DAGM Symposium (DAGM'06)*, Berlin, Germany, 2006, *LNCS 4174*, pp. 729–738. Springer.

[11] M. Magnor, K. Hildebrand, A. Lintu, and A. J. Hanson. Reflection nebula visualization. In C. Silva, E. Gröller, and H. Rushmeier, eds., *IEEE Visualization 2005, VIS 2005*, Minneapolis, USA, 2005, pp. 255–262. IEEE.

[12] V. Scholz and M. Magnor. Multi-view video capture of garment motion. In P. H. de With, C. Varekamp, D. S. Farin, and Y. Morvan, eds., *Content Generation and Coding for 3D-Television*, Eindhoven, Netherlands, 2006, pp. 1–4. Technische Universiteit Eindhoven.

[13] V. Scholz and M. Magnor. Texture replacement of garments in monocular video sequences. In D. W. Fellner, S. N. Spencer, Y. Chrysanthou, D. Cohen-Or, T. Akenine-Möller, and W. Heidrich, eds., *Rendering Techniques 2006, Eurographics Symposium on Rendering*, Nicosia, Cyprus, June 2006, pp. 305–312. Eurographics.

[14] V. Scholz, T. Stich, M. Keckeisen, M. Wacker, and M. Magnor. Garment motion capture using color-coded patterns. *Computer Graphics forum*, 24(3):439–448, September 2005. Conference Issue: 26th annual Conference Eurographics 2005, Dublin, Ireland, August 29th - September 2nd, 2005.

[15] T. Stich and M. Magnor. Learning flames. In G. Greiner, J. Hornegger, H. Niemann, and M. Stamminger, eds., *Vision, Modeling, and Visualization 2005 (VMV'05)*, Erlangen, Germany, 2005, pp. 60–65. Akademische Verlagsgesellschaft Aka.

[16] C. Theobalt, M. Magnor, and H.-P. Seidel. 3D image analysis and synthesis at MPI informatik. In D. Fellner, ed., *Vision, Video, and Graphics 2005*, Edinburgh, UK, 2005, pp. 85–92. Eurographics.

Theses

[1] S. El-Abed. A computer graphics-based analysis-by-synthesis approach for the joint recovery of the gas and dust distributions in bipolar planetary nebulae from visible and radio wavelength observations. Bachelor thesis, Universität des Saarlandes, September 2006.

[2] K. Hildebrand. Rendering and reconstruction of astronomical objects. Masters thesis, Universität Weimar, September 2005.

[3] C. Linz. Pde-based surface reconstruction from multiple views using a surfel model. Masters thesis, Universität des Saarlandes, September 2005.

[4] M. Magnor. *Visuelle Ästhetik vs. Visual Computing.* Habilitation thesis, Universität des Saarlandes, January 2005.

[5] F. Recktenwald. Sampling and interpolation of the plenoptic function from sparse data. Masters thesis, Universität des Saarlandes, January 2006.

[6] M. Schaef. Rekonstruktion von bewegtem Stoff mittels direkter Suchverfahren. Bachelor thesis, Universität des Saarlandes, January 2005.

Technical reports

[1] C. Theobalt, N. Ahmed, E. de Aguiar, G. Ziegler, H. Lensch, M. Magnor, and H.-P. Seidel. Joint motion and reflectance capture for creating relightable 3d videos. Research Report MPI-I-2005-4-004, Max-Planck-Institut fuer Informatik, Saarbrücken, Germany, April 2005.

Part IV

Index

Index