2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671730

COVID-19++: A Citation-Aware Covid-19 Dataset for the Analysis of Research Dynamics

Lukas Galke
ZBW
Kiel, Germany
l.galke
@zbw.eu

Eva Seidlmayer
ZB MED
Cologne, Germany
seidlmayer
@zbmed.de

Gavin Lüdemann
ZBW
Kiel, Germany
stu205587
@mail.uni-kiel.de

Lisa Langnickel
ZB MED
Cologne, Germany
langnickel
@zbmed.de

Tetyana Melnychuk
Kiel University
Germany
melnychuk
@bwl.uni-kiel.de

Konrad U. Förstner
ZB MED—Information Centre for Life Sciences
Cologne, Germany
foerstner@zbmed.de

Klaus Tochtermann
ZBW—Leibniz Information Centre for Economics
Kiel and Hamburg, Germany
k.tochtermann@zbw.eu

Carsten Schultz
Kiel University
Germany
schultz@bwl.uni-kiel.de

***Abstract*—COVID-19 research datasets are crucial for analyzing research dynamics. Most collections of COVID-19 research items do not to include cited works and do not have annotations from a controlled vocabulary. Starting with ZB MED KE data on COVID-19, which comprises CORD-19, we assemble a new dataset that includes cited work and MeSH annotations for all records. Furthermore, we conduct experiments on the analysis of research dynamics, in which we investigate predicting links in a co-annotation graph created on the basis of the new dataset. Surprisingly, we find that simple heuristic methods are better at predicting future links than more sophisticated approaches such as graph neural networks.**

***Index Terms*—COVID-19, SARS-CoV-2, coronavirus, dataset, link prediction, network analysis, research dynamics, graph neural networks, covid-19 dataset, research collection**

I. Introduction

We strive to assemble a collection of COVID-19 research articles that includes bibliographic metadata from cited works and subject labels from a controlled vocabulary, i. e., concepts. Such a holistic collection of COVID-19 research would facilitate the analysis of research dynamics, and, in particular, how science responds to an urgent global crisis. To assemble the dataset, we combine bibliographic metadata from different sources, which is challenging because they come with different identifiers, formats, and use different controlled vocabularies (or none). Furthermore, the research in this area is still ongoing. Thousands of new publications appear every month. Thus, the assembly procedure is as important as the data itself.

Previous collections of COVID-19 research such as CORD19 [1] or the COVID-19 subset of the ZB MED Knowledge Environment[1] do not include cited works. In this work, we combine bibiographic metadata from KE with citation data from CrossRef and add the cited works to the dataset. We further harmonize author data, fetch journal identifiers, and annotate all papers with Medical Subject Headings[2] (in short: MeSH terms), where those are missing.

After assembling the dataset, we conduct experiments on link prediction in the co-annotation graph to analyze to what extent the research directions, expressed as the co-occurrence of MeSH terms, can be known ahead of time. In practice, link prediction in co-annotation graphs could be used for recommending promising research directions to researchers. Such applications are only possible because of the expanded view of our newly assembled dataset.

In summary, our contributions are:

- We provide a new dataset of COVID-19 publication data.
- The dataset contains COVID-19 research papers along with first-order cited work.
- We use ConceptMapper [2] to generate MeSH annotations, whenever those annotations are not present.
- We conduct experiments on link prediction between concepts from the newly created dataset.
- We describe the procedure for assembling the dataset and provide the code for keeping the data collection up-to-date.

II. Related Data Collections

We describe existing collections of COVID-19 research articles that are relevant to the dataset introduced in this work.

a) CORD-19: COVID-19 Open Research Dataset[3] [1]. CORD-19 is a free and open dataset of research articles on COVID-19. It is maintained by the Semantic Scholar team at the Allen Institute for AI in collaboration with leading research groups. As of Aug 9, 2021, The dataset covers more than 280,000 scholarly articles.

b) CrossRef: CrossRef has released a 65GB data file[4] to support COVID-19 research. The file contains 112M metadata records. These records are, however, not limited to COVID-19

This work was supported by BMBF under grant numbers 01PU17013A, 01PU17013B, 01PU17013C. The first five authors have contributed almost equally.

[1] https://www.livivo.de/covid19

[2] https://www.nlm.nih.gov/mesh/meshhome.html

[3] https://www.semanticscholar.org/cord19

[4] https://www.crossref.org/blog/free-public-data-file-of-112-million-crossref-records/

research but contain everything that is registered with CrossRef until March 2020. Apart from this static collection, CrossRef also provides an API to retrieve up-to-date metadata, which we have used for the enrichment of our dataset.

c) ZB MED KE: The ZB MED Knowledge Environment (KE)[5] provides bibliographic metadata from the holdings of ZB MED, as well as important data sources such as MEDLINE, AGRIS, AGRICOLA (National Agricultural Library), National Library of Medicine and others, also including scientific publishers such as Thieme and Karger, as well as a selection of data sources from the open access database BASE. With this, KE has combined information on medicine and health on the one hand, and nutritional, environmental and agricultural sciences on the other, and fully leverages the synergy effects through the integration of three thesauri: MeSH, Agrovoc, and UMTHES to use. In total, the KE holds more than 70M items from more than 50 scientific databases The KE also lists MeSH terms, where available. The CORD-19 dataset has been migrated into KE and is extended and curated manually. As of Aug 9, 2021, the COVID-19 subset of the KE currently holds about 50,000 records of scholarly articles on COVID-19.

d) Summary: In summary, the KE includes CORD-19 plus manual curation. However, citation data are missing in the considered collections of COVID-19 research. Moreover, MeSH annotations are not prevalent or not complete. Since citation data are essential for the analysis of research dynamics, we extend the KE with citation data and metadata of cited works, while also adding missing concept annotations.

III. The COVID-19++ Dataset

Below, we describe the procedure of assembling our new COVID-19++ dataset including our automated annotation and data cleaning procedures.

a) Primary Data Sources: As primary data sources, we use the ZB MED KE subset of COVID-19 along with further preprint sources. We harvested the *COVID-19 collection* of the KE which includes about 50k entries from CORD-19 as well as relevant preprints from bioRxiv and medRxiv. We harvested only scientific articles with publication dates between January 1, 2020 to December 31, 2020. We requested DOI, title, publication date, and MeSH annotations if available.

Furthermore, we retrieved preprints from five different preprint servers via the API of *preVIEW COVID-19*, which is a ZB MED hosted, publicly available semantic search engine for COVID-19 preprints [3]. It contains articles from medRxiv, bioRxiv, preprints.org, chemRxiv and arXiv. These are either selected via specific APIs, e.g., bio- and medRxiv provide a shared COVID-19 API, or via dedicated queries that search for key terms such as ”corona”, ”COVID-19” or ”SARS-CoV-2” and filter the articles by date. The exact search queries can be found under https://github.com/ZBMED/preVIEW-COVID19.

b) Fetching data for cited works: We enrich the data of ZB MED KE and preprints with citation information from the CrossRef bibliographic database. On the basis of the DOIs, we have retrieved all research items from CrossRef that were cited by at least one publication of the original dataset related to COVID-19 subjects. After retrieving the DOIs of the cited works, we consulted ZB MED KE to retrieve their metadata records including title, publication year, and abstracts. We use the abstract for automated MeSH annotation.

c) Annotation with MeSH terms: Preprints and cited articles from CrossRef do not come with MeSH annotations. For the reviewed articles, we first look up whether the DOI can be found in ZB MED KE such that we can fetch the annotations from there. However, there is no manual indexing for non-peer-reviewed articles. Therefore, we use the dictionary lookup-based annotation tool ConceptMapper [2], which is based on the Apache Unstructured Information Management Architecture (UIMA) [4] environment. ConceptMapper is a highly configurable tool that allows for advanced string matching in a reasonable amount of time. We used the implementation of Funk et al. [5] which is available on GitHub[6].

TABLE I
Comparison of ConceptMapper and MeshOnDemand.
Sample-based precision (P), recall (R), and F_1-score times 100.

#MeSH Terms	#Docs	Method	P	R	F_1
3,722 (cleaned)	1001	ConceptMapper	46.69	34.00	35.43
3,722 (cleaned)	1001	MeshOnDemand	59.77	48.87	49.59
567 (diseases)	550	ConceptMapper	62.45	42.41	38.71
567 (diseases)	550	MeshOnDemand	78.64	61.61	57.97

To assess the quality of the produced annotations, we have evaluated ConceptMapper on a subset of 1001 articles, for which we do have the ground truth annotations. We use both tools to make predictions for these articles and eveluate them with sample-based precision, recall, and F_1-score. We have excluded 59 generic MeSH terms (e. g., “Humans”, “Mice”, and “Methods”) from this evaluation. We find that ConceptMapper achieves a document-averaged F_1-score of 35.43 using a vocabulary of 3,722 distinct MeSH terms. This is remarkable as ConceptMapper achieves 71.4% (66.8% for diseases) of the performance of the official MeSH recommendation tool.

We list the MeshOnDemand scores for comparison, even though the system has had access to our test documents: we observed frequently that the original document is listed as the top-1 similar article. This means that the scores of MeshOnDemand are likely higher than they would be with truly unseen documents, as we are facing. In contrast, we can expect that the string-matching ConceptMapper attains a similar performance on the new documents. We retain the information whether the labels were generated automatically or come from manual indexing. Thus, future work could use a different annotation method, e. g., a multi-layer perceptron text classifier [6]. For now, we prefer to use more explainable annotations from ConceptMapper.

d) Harmonizing author data: Unfortunately, the format for author data differs between different data sources

[5] https://www.livivo.de/

[6] https://github.com/UCDenver-ccp/ccp-nlp-pipelines

(`Lastname, Firstname` vs. `Firstname Lastname`). To unify author names without relying on heuristics, we fetch all author names from CrossRef by using the articles' DOIs. We also retrieve the author's unique ORCID from CrossRef, when available.

e) Data Merging and Cleaning: Finally, we merge the data from different sources and conduct several data cleaning steps. Those are (in-order):

1) Drop any papers, in which the title or the date is missing.
2) Drop duplicate papers while giving preference to the ZB MED KE metadata over preprints over referenced works.
3) Remove qualifiers from MeSH terms and drop duplicates.
4) Remove annotations of previously removed papers.
5) Remove MeSH terms that occur in less than 20 papers.
6) Ensure that each paper has at least one annotation within the retained MeSH terms and discard others.
7) Remove duplicate authors.
8) Ensure that each paper has at least one author.
9) Ensure that each paper has at least one annotation, again.
10) Discard any authors that are listed on less than 2 papers.

By step 8, we ensure that each record is complete and remove those publications for which author data were not available. In step 10, we reduce the total dataset size without removing any articles that might carry important information.

TABLE II
COVID-19++ DATASET CHARACTERISTICS

Characteristic	Value
# Original Publications	47,242
# Added Preprints	11,154
# Cited Publications	279,027
# Total Publications	337,423
# Citations	906,110
# Authorships	1,421,320
# Unique Authors	367,743
# Concept Annotations	5,255,479
# Unique Concepts	13,253
Avg. #Citations per Publ.	2.7
Avg. #Author per Publ.	4.2
Avg. #Concepts per Publ.	15.6

f) Resulting Dataset Characteristics: The resulting dataset, which we call COVID19++, consists of 337,423 publications (and preprints). We have over 900k citations, from which both the citing and the cited work are within our dataset. The dataset comprises over 5M annotations with concepts from a controlled vocabulary, i. e., MeSH terms. Table II summarizes the basic characteristics of the resulting dataset. On average, we have 2.7 citations per publication, 4.2 authors per publication, and 15.6 concepts per publication. Figure 1 shows the count of publications by date within our COVID-19++ dataset. We see that by adding first-order cited research items, we have obtained a dataset that ranges back more than half a century and provides a solid foundation for the analysis of short and long-term research dynamics.

IV. Link Prediction Experiments

The advantage of COVID-19++ is that it holds an expanded view on COVID-19 research that also includes cited previous work. This facilitates the analysis of research dynamics in response to the COVID-19 pandemic in a more holistic manner. As an exemplary application of our dataset, we tackle the question: *To what degree were research directions foreseeable from prior data alone?*

We investigate this question by looking at pairs of concepts and the point in time at which a paper has been annotated with both of them for the first time. If the same paper has been annotated with two concepts, then we assume that the concepts are *linked* in some way. Hence, we first transform the data into a *co-occurrence* graph with a fixed vertex set and dynamic edges. Vertices are concepts from the controlled vocabulary, and edges are papers that link those concepts. Then, we predict future links between concepts, i. e., estimate the likelihood that a research paper connecting any two concepts will appear.

A. Problem Statement

We denote the set of concepts (or, more abstract, the set of co-occurrants) as $\mathbb{C}$. We say that any two elements $u, v \in \mathbb{C}$ co-occur at time t if t is the earliest time at which a single paper has been annotated with both concepts u, v. We construct an undirected graph $\mathcal{G} = (V, E)$ with the vertex set $\mathcal{V}$ being the set $\mathbb{C}$ of co-occurrants. We insert an edge $(u, v) \in E$ when nodes u and v cooccur. For each edge (u, v), we store the earliest time of co-occurrence. All co-occurrants associated with less than $k_{\min}$ or more than $k_{\max}$ papers are removed. Finally, we consider only the largest connected component.

The aim is to learn a function $f : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ that estimates the score of any two nodes getting connected in the future. Thus, the score $f(u, v)$ should be high for future links and low otherwise. We evaluate f on all candidate pairs from the test set to produce a ranked list, which we compare against true future links via ranking metrics.

B. Link Prediction Methods

In the following, we first recapture heuristic link prediction methods, before we describe methods that rely on node embeddings, and, subsequently, graph autoencoders and SEAL.

a) Heuristic Link Prediction Methods: Common heuristic methods for link prediction are Preferential Attachment (PA) [7], Common Neighbors (CN) [8], and Adamic/Adar (AA) [9]. Preferential Attachment multiplies the degree of both involved nodes to determine a score: $\mathrm{PA}(u, v) = \deg(u) \cdot \deg(v)$ The argument for using this score as a link prediction measure is that well-connected nodes are more likely to gain links. PA does not take locality into account.

Common Neighbors, in contrast, is entirely based on locality. It counts how many neighbors two nodes have in common: $\mathrm{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|$, where $\Gamma(v)$ is the set of neighbors of v. To count common neighbors efficiently, we take the second power of the adjacency matrix $\boldsymbol{A}^2$, similar to Galke et al. (2018) [10].

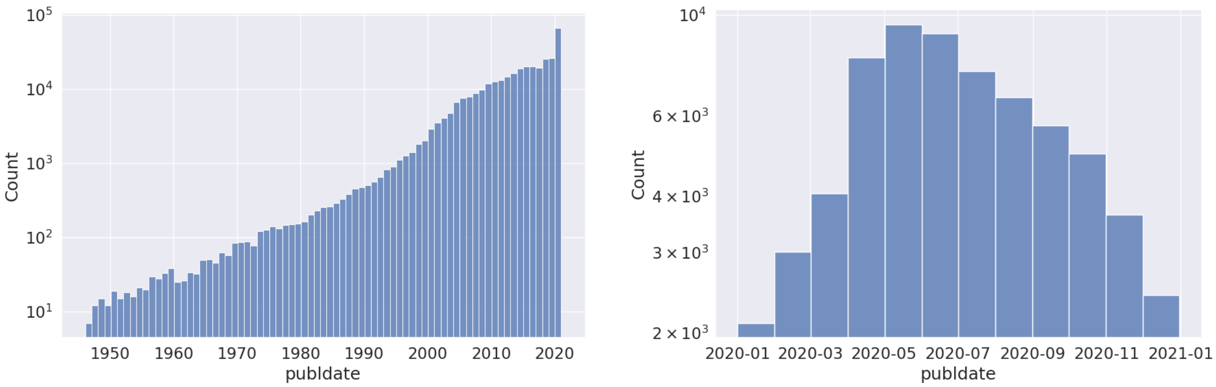

Fig. 1. Non-cumulative count of publications (log-scale) by publication date in our COVID-19++ dataset *(Left)*. Non-cumulative count of publications (log-scale) since 2020 by publication date in our COVID-19++ dataset *(Right)*

Adamic/Adar adjusts the weights of common neighbors according to their degree, giving nodes with a higher degree less influence: $\mathrm{AA}(u,v) = \Sigma_{w\in\Gamma(u)\cap\Gamma(v)}\frac{1}{\log(\deg(w))}$.

b) Node Embeddings: DeepWalk [11] and node2vec [12] are established methods to learn node representations in a graph. The representations are learned by sampling random walks through the graph and predicting neighboring vertices along the path, while updating the node embeddings according to the error signal of the prediction. Those learned representations may be used as input for a link prediction task. Common choices for the final predictor module are the inner product of the representations, or, a parametrized affine transformation of their Hadamard product.

c) Graph Autoencoder (GAE) [13]: Graph neural networks (GNNs) can be used to reconstruct the adjacency matrix as in (variational) graph autoencoders [13]. We use a GAE architecture with an encoder composed of a learnable node embedding layer, followed by graph convolution layers. Every layer has twice the output dimension of the one below it, creating a funnel shape. The latent dimension is then equal to the output dimension of the final layer. Again, we evaluate two different decoder modules: one is an inner product decoder, the other applies a linear layer on top of the Hadamard product of the paired representations. Both decoders use a sigmoid activation on the output.

d) SEAL [14]: Finally, we apply a state-of-the-art link prediction method that is based on graph convolution, namely, SEAL. SEAL adds several features to the graph autoencoder framework with link prediction as the key goal. The idea is to operate on a link-centric subgraph that is labeled by a breadth-first strategy to increase the GNN's expressive power. Thus, the GNN operates on the surrounding subgraph of each *potential* link. The underlying graph neural network is DGCNN [15] For the output, SEAL concatenates all intermediate layers' representations as in jumping knowledge networks [16], and then summarizes the graph across nodes via sort pooling (we used top-60% as suggested by the original work). Finally, a 1d-convolution with max-pooling before a multilayer perceptron with dropout 0.5 yields the binary classification result.

C. *Experimental Procedure*

We use the COVID-19++ dataset as described in Section III and create the concept graph as described in Section IV-A.

We follow [17] for a rigorous link prediction evaluation. First, it is crucial to use a chronological train/test split such that we predict *future* links, which is more challenging than missing links. Second, negative subsampling of the test set may bias the evaluation. In our work, we compare balanced 50-50 test sets, which are subsampled, against complete (non-subsampled) test sets created on the basis of geodesic distance, i. e., the length of the shortest path between two nodes.

We now formally define our chronological train-test split. Given t_0, t_1 and t_2, G is split into two graphs $G_{\text{train}} = (E_{\text{train}}, V)$ and $G_{\text{test}} = (E_{\text{test}}, V))$ s.t. if u, v co-occur at time t, then $(u,v) \in E_{\text{train}} \iff t_0 \leq t < t_1$ and $(u,v) \in E_{\text{test}} \iff t_1 \leq t < t_2$. This way, we ensure that E_{train} and E_{test} are disjoint.

We created two different train/test splits, the *large* and *small training split*. The large training split was created with $t_0 := 01/01/1950, t_1 := 07/01/2020$ and $t_2 := 01/01/2021$. Therefore, the task is to predict the second half of 2020 using the complete set of data from the past. The training set contains $4,368,979$ edges and the test set contains $222,780$ edges for a ratio of about $20:1$. The small train set, on the other hand, with $t_0 := 01/01/2020$ leads to $378,932$ training edges. We have visualized the degree distributions of both training set configurations and the test set in Figure 2. Note that the degree distribution of 2020/01-2020/06 is more similar to the test set than the 1950-2020/06 training set.

We use these splits to evaluate a method's performance at the task described in Section IV-A by training the respective model, if applicable, and predicting links on the training set and then evaluating these predictions on the test set.

We trained our GAE models using the following hyperparameters: an initial 64-dimensional node embedding is fed into two graph convolution layers with 32 latent dimensions. For each batch, we sampled 64 subgraphs using 2-hop random walk sampling. The model was trained for 12800 and 6400

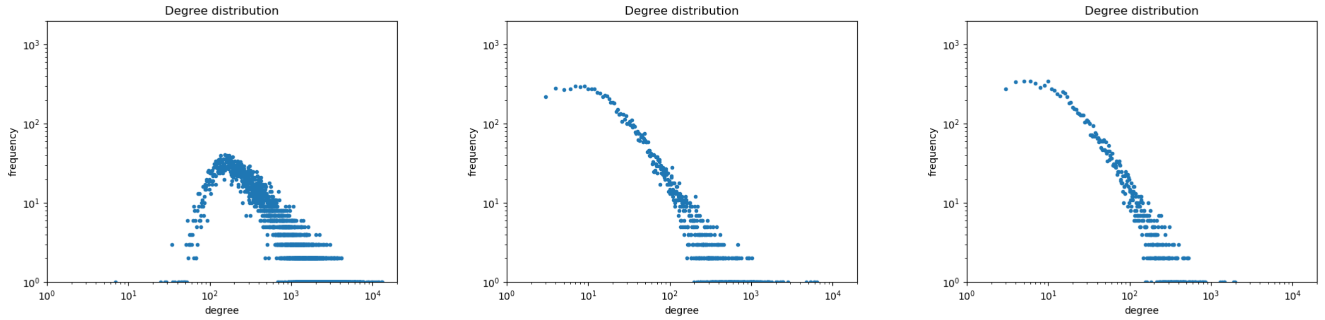

Fig. 2. Degree distributions of co-annotation graphs: large train set 1950–2020/06 *(Left)*, small train set 2020/01–2020/06 *(Center)*, and test set 2020/07–2020/12 *(Right)*. Axes are on log-scale with frequency on y-axis and degree on x-axis. The small train set seems to be more representative of the test set.

iterations on the large and small training split, respectively, using Adam [18] with a learning rate of 10^{-3} and weight decay of $5 \cdot 10^{-4}$. We use early stopping for the GAE models: After each epoch (128 iterations), we evaluate the whole training set are evaluated and we store the best model. After training, the best model is restored and used for evaluation on the test set.

The test graphs were too large in terms of $\mathcal{O}(n^2)$ node pairs to feasibly evaluate predictions for every possible node pair. Therefore, we employ two methods of selecting test data: geodesic distance and negative subsampling.

a) Candidate pairs on the basis of geodesic distance: Geodesic distance $d(v, u)$ between nodes v and u is defined as the length of the shortest path between the two nodes. In this approach, we compose the test set by selecting all pairs of nodes v, u with $d(v, u) = k$ for some k [17]. We use $k = 2$ because $k = 3$ would lead to only a small number of candidate pairs. The main disadvantage of this approach is that the ratio of negative and positive samples depends on the specific graph and even on the train/test split, making comparisons between scores on different graphs difficult. On the other hand, neither positive nor negative edges are subsampled, making this a deterministic sampling method (given a fixed train/test split). The sets can be very large, especially if the underlying graph is dense; for $k = 2$ we get $83,414,428$ and $28,038,381$ node pairs on the large and small training split, respectively.

b) Negative subsampling: We consider all positive edges in the test graph, as well as an equal number of randomly selected edges that are neither in the training graph nor in the test graph. This approach might lead to overly optimistic results [17]. We still include it in our comparison to determine the difference to a more profound link prediction evaluation on the basis of geodesic distance.

D. Evaluation Measures

We follow [17] and use average precision and area under the receiver operating characteristic as performance measures on previously unseen candidate edges.

Average Precision (AP) is defined as $\sum_{n=1}^{N} P_n(R_n \cdot R_{n-1})$, where P_n and R_n are precision and recall of the top n results within the full result set of size N. In effect, AP is the mean precision at every possible threshold, weighted by the gain in recall relative to the next higher threshold. This weighting is more accurate than interpolating the PR-curve [19]. For random guessing and binary labels, the expected AP is equal to $\frac{|C_p|}{N}$, where C_p is the set of positives.

The area under the receiver operating characteristic curve (AUROC) is defined as the (linearly interpolated) area under the ROC curve, which is obtained by plotting the true positive rate against the false positive rate for every threshold. For random guessing, the expected AUROC is 0.5.

E. Results

We measure the performance of the methods described in Section IV-B applied to the splits described in Section IV-C. Table III summarizes our results in terms of area under the ROC curve and average precision for the *large training split*, evaluated using either negative subsampling (50-50) or geodesic distance sampling with $k = 2$ for the compared link prediction methods, as well as a random predictor with $p = 0.5$. Table IV shows our results for the *small training split*.

TABLE III
LINK PREDICTION RESULTS WITH LARGE TRAINING SPLIT (SINCE 1950). AUROC/AP SCORES OF THE COMPARED METHODS ON CANDIDATE PAIRS FROM THE SECOND HALF OF 2020. HIGHER IS BETTER.

Method	50-50	gd=2
Random	50.01 / 50.02	49.92 / 0.25
Preferential Attachment	87.06 / 86.04	87.12 / 2.53
Common Neighbors	85.61 / 84.71	85.58 / 2.24
Adamic/Adar	**89.75 / 88.62**	**89.36 / 3.07**
Node2vec (InnerProduct)	49.99 / 49.94	49.98 / 0.25
Node2vec (Hadamard+Linear)	49.88 / 49.78	49.88 / 0.25
GAE (InnerProduct)	57.29 / 54.07	57.32 / 0.29
GAE (Hadamard+Linear)	62.25 / 62.77	62.23 / 1.34
SEAL	89.26 / 32.90	89.25 / 2.92

We observe that even simple methods such as Adamic/Adar can reliably predict links in the 50-50 split of the candidate test edges. Still, our evaluation on the basis of geodesic distance reveals that this does not imply that COVID-19 research directions are easily predictable. This is particularly important because a geodesic distance evaluation corresponds to real-world usage of a link prediction system. On the geodesic distance test set with the small training set, we find that the

TABLE IV
LINK PREDICTION RESULTS WITH SMALL TRAINING SPLIT (SINCE 2020/01). AUROC/AP SCORES OF THE COMPARED METHODS ON CANDIDATE PAIRS FROM THE SECOND HALF OF 2020. HIGHER IS BETTER.

Method	50-50	gd=2
Random	49.96 / 49.88	50.03 / 0.65
Preferential Attachment	87.93 / 88.82	**87.61** / 8.18
Common Neighbors	88.34 / **89.62**	85.48 / **8.48**
Adamic/Adar	86.20 / 86.90	81.56 / 6.81
Node2vec (InnerProduct)	50.05 / 49.98	49.97 / 0.65
Node2vec (Hadamard+Linear)	50.03 / 50.04	50.04 / 0.65
GAE (InnerProduct)	80.33 / 76.07	64.95 / 0.94
GAE (Hadamard+Linear)	73.14 / 68.26	50.56 / 0.66
SEAL	**90.99** / 86.50	86.31 / 6.15

heuristic methods perform best with scores between 6.81 and 8.48. Only SEAL comes close with 6.15, while GAE and node2vec hardly exceed chance performance.

V. DISCUSSION

We have assembled a new dataset of COVID-19 that includes citation data and concept annotations. We have used this dataset for experiments on link prediction in a co-annotation graph with a chronological train-test split. To interprete our results, the AUROC scores for predicting concept links look deceptively strong because the area under the ROC curve does not reflect the class imbalance. Therefore, we have included another evaluation strategy consisting of the average precision on a non-subsampled test set, which better resembles real-world applications [17]. This evaluation reveals that the prediction of COVID-19 research direction is challenging and it is not easy to foresee research directions. Still, heuristic methods do better than chance and, surprisingly, even outperform state-of-the-art deep learning methods.

A potential bias of our dataset assembly is the inclusion of preprints, as preprints do not guarantee acceptance and publication. However, between 17% and 30% of total COVID-19 research papers published in 2020 were preprints [20]. Thus, preprints play a major role in COVID-19 research and we argue that the potential gains of including preprints outweigh the dangers of bias. Only 15 preprints and, on the other hand, 24 journal articles had been withdrawn or retracted by December, 2020 [20]. Furthermore, publishing a preprint is associated with more citations for the final peer-reviewed article [21].

A different concern is that the link prediction would reinforce the status quo. However, we made sure that the test set of our evaluation only contains concept pairs that did not appear anytime before. As such, a status quo prediction is *not* regarded as true positive during evaluation and can be easily filtered out from the predictions in practice.

COVID-19 is still an active area of research and new publications are appearing every month. Thus, we make the scripts for harvesting, data cleaning, and preprocessing available for reuse. We hope that future research can build upon our work to investigate research dynamics or to further develop link prediction techniques that can be used to recommend combinations of concepts as candidate research topics.

VI. CONCLUSION AND FUTURE WORK

This work introduces a new collection of COVID-19 research articles with citations and MeSH labels. We use this dataset to conduct experiments on link prediction, which might be used to recommend new combinations of subjects to researchers. We make the dataset available at https://doi.org/10.5281/zenodo.5531084. We further provide the scripts for harvesting and curating the data (https://github.com/Q-AKTIV/covid19-harvesting-tools), with which new versions of this dataset can be generated. In future work, we plan to use this dataset to analyze the temporal evolution of research topics.

REFERENCES

[1] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. M. Kinney, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. A. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "Cord-19: The covid-19 open research dataset," *ArXiv*, 2020.

[2] M. A. Tanenblatt, A. Coden, and I. L. Sominsky, "The conceptmapper approach to named entity recognition." in *LREC*. Citeseer, 2010.

[3] L. Langnickel, R. Baum, J. Darms, S. Madan, and J. Fluck, "COVID-19 preVIEW: Semantic search to explore COVID-19 research preprints," in *Studies in Health Technology and Informatics*. IOS Press, 2021. [Online]. Available: https://ebooks.iospress.nl/doi/10.3233/SHTI210124

[4] D. Ferrucci, A. Lally, K. Verspoor, and E. Nyberg, "Unstructured information management architecture (uima) version 1.0," 2008.

[5] C. S. Funk, W. A. B. Jr., B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter, and K. Verspoor, "Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters," *BMC Bioinform.*, vol. 15, 2014.

[6] L. Galke and A. Scherp, "Forget me not: A gentle reminder to mind the simple multi-layer perceptron baseline for text classification," 2021.

[7] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical mechanics and its applications*, vol. 311, no. 3-4, 2002.

[8] M. E. Newman, "Clustering and preferential attachment in growing networks," *Physical review E*, vol. 64, no. 2, 2001.

[9] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Soc. Networks*, vol. 25, no. 3, 2003.

[10] L. Galke, F. Mai, I. Vagliano, and A. Scherp, "Multi-modal adversarial autoencoders for recommendations of citations and subject labels," in *UMAP*. ACM, 2018.

[11] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: online learning of social representations," in *KDD*. ACM, 2014.

[12] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *KDD*. ACM, 2016.

[13] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *CoRR*, vol. abs/1611.07308, 2016.

[14] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *NeurIPS*, 2018.

[15] A. V. Phan, M. Le Nguyen, Y. L. H. Nguyen, and L. T. Bui, "Dgcnn: A convolutional neural network over large-scale labeled graphs," *Neural Networks*, vol. 108, 2018.

[16] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in *International Conference on Machine Learning*. PMLR, 2018.

[17] Y. Yang, R. N. Lichtenwalter, and N. V. Chawla, "Evaluating link prediction methods," *Knowl. Inf. Syst.*, vol. 45, no. 3, 2015. [Online]. Available: https://doi.org/10.1007/s10115-014-0789-0

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[19] P. A. Flach and M. Kull, "Precision-recall-gain curves: PR analysis done right," in *NIPS*, 2015.

[20] H. Else, "How a torrent of covid science changed research publishing-in seven charts." *Nature*, 2020.

[21] D. Y. Fu and J. J. Hughey, "Meta-research: Releasing a preprint is associated with more attention and citations for the peer-reviewed article," *Elife*, vol. 8, 2019.