Codd's World: Topics and their Evolution in the Database Community Publication Graph

Rutuja Shivraj Pawar, Sepideh Sobhgol, Gabriel Campero Durand, Marcus Pinnecke, David Broneske and Gunter Saake

Otto-von-Guericke University Magdeburg Germany

rutuja.pawar@ovgu.de, sepideh.sobhgol@st.ovgu.de, campero@ovgu.de, pinnecke@ovgu.de, david.broneske@ovgu.de, saake@ovgu.de

ABSTRACT

Scholarly network analysis is the study of a scientific research network aiming to discover meaningful insights and making data-driven research decisions. Analyzing such networks has become increasingly challenging, due to the amount of scientific research that is added every day. Furthermore, online resources often include information from other online sources (e.g., academic social platforms), enabling to study networks on a larger and more complex scope. In this paper, we present a study on a specific research network: The (relational) database community publication graph, that we call *Codd's World*; a transitive closure over citations from the foundational work of E.F. Codd. We specifically analyze the topics of the published papers, the relevance of authors and papers, and how this relates to raw publication counts. Among our findings, we show that topic modeling can be a useful entry point for scholarly network analysis.

Keywords

Data Analysis, Topic Modeling, Database Publication Network, Science of Science

1. INTRODUCTION

Rapid advancements in science and research leads to enormous amounts of digital scholarly data being produced and collected every day [1]. This scholarly data can be in the form of scientific publications, books, teaching materials, and many other scholarly sources of information made available on the Web. Apart from the *volume* of scholarly data, there is a meaningful variety of *connections* in this data. For instance, papers are connected through citations, authors are connected in networks of collaboration, and there are many other embedded relationships. As a consequence, in the world of research, understanding relevant work and its scientific impact has become more and more challenging. One approach that allows modeling the relevance of papers by leveraging the networks in which they are embedded is *scholarly network analysis (SNA)*.

SNA proposes to study the underlying structure of a scholarly network, helping the research community to implement data-driven decisions. SNA compromises of at least seven points of interest: (i) authors collaborating on papers (*co-authorship* networks), (ii) documents referencing each other (*citation* networks), (iii) documents cited together (*co-citation* networks), (iv) documents that cite other documents in a similar manner (*bibliographical coupling*), (v) content clusters in document sets (*topic* networks), (vi) word clusters occurring together (*co-words*), and (vii) any diversity of types and relationships (*heterogeneous networks*) [2]. Some example SNA studies to mine knowledge from these scholarly networks include analyzing the citation relationships to evaluate the impact of a given paper or an author [3] or studying co-author behavior to identify the scientific community distribution [4]. Furthermore, *topic network analysis (TNA)* can be used to extract the underlying topics from a corpus in terms of word distribution and also the affinity of each document to a topic. Discovering the topics unveils the underlying structure of the data and thus better serves as a first step towards SNA. Furthermore, the visualization of the extracted topics can lead to discovering topic evolution over time [5–7].

Main Contributions: In this paper we undertake an SNA study over a sub-network from the DBLP dataset of computer science publications [8] with TNA as our starting point towards understanding this complex network structure. Specifically, we select a network that corresponds to the (relational) database influence community. We call this specific network *Codd's World*. Our main contributions can be summarized as follows:

(i) Unlike previous studies about the database community, which have restricted the community to encompass papers appearing in top database venues, like VLDB, ICDE, EDBT and SIGMOD [9]; in this paper we identify the community by using the foundational work of E.F. Codd [10] as a starting point, and collect all papers transitively related to this work through citation relationships. As a result, we expect our work to show a more diverse view of the database community, spanning influences beyond the top database venues.

(ii) Unlike previous studies about the database community, in our work, we propose to consider topics as an important dimension to understand the underlying structure of the publication network; and how it can be a first step for visualizing the content and discovering meaningful trends. Hence, this paper presents a detailed description of how unveiling the research topics was utilized as the first step towards SNA on the database publication graph. We complement this approach by analyzing relevance and the role of self-citations.

The rest of the paper is structured as follows: The relevant basic background is presented in Sec. 2. The SNA carried out with the

31st GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 11.06.2019 - 14.06.2019, Saarburg, Germany.

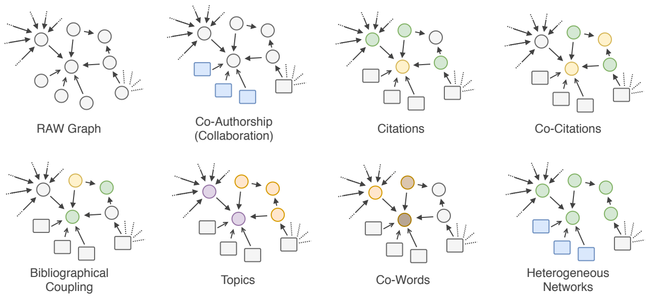

Figure 1: Overview of SNA on raw graphs: co authorship/collaboration, citations, co-citations, bibliographical coupling, topics, co-words, and heterogeneity.

results answering the formulated research questions is detailed in Section 3. Section 4 concludes the paper discussing the next steps in this research direction.

2. BACKGROUND

In this section we provide background on concepts and terms used in this work. We start with SNA (Section 2.1), continue with topic models (Section 2.2), relevance ranking (Section 2.3) and end with detection of self citations (Section 2.4).

2.1 Scholarly Network Analysis

Figure 1 shows various aspects of SNA depicting the seven types of networks usually considered for SNA on RAW graphs. Further elaborating on their functionality, Citation and co-citation network analysis is used to find relationships between cited papers and a set of papers which cite those papers. Moreover, citation analysis can be employed in community detection which is one of the fundamental tasks of network analysis [11]. Bibliographic coupling also employs citation analysis to link documents which reference a common third cited work in their bibliography [12]. Co-authorship network analysis can be used to find scientific collaboration between authors [13] depicting how individual scientific ideas of authors can get together through collaboration to cause an explosion of scientific findings. Through co-word analysis [14], it is possible to identify the relationships between subjects in the specific field of research through finding the co-occurrence of keywords which helps to examine the development of science in specific areas. Discovering citation evolution over time or future citation through coupling co-authorship and citation networks can be considered as a task of heterogeneous network analysis [15].

2.2 Topic Models

Topic Models (TMs) are unsupervised learning models that, when given a set of documents as input, learn the underlying topics in terms of word distribution and also the affinity of each document towards a topic [5]. In SNA, TMs are useful to discover the current research topics as well as its relevant publications and to identify topic trends through time. Among the many prevalent topic modeling techniques, the following have seen a growing utilization in various applications.

Latent Dirichlet Allocation (LDA) LDA is a probabilistic generative topic model used for modeling of discrete data collections through learning the relationships between words, topics and documents [16, 17]. LDA views each document as a mixture of topics and through its processing, LDA assigns documents with the modeled topics. The word distribution in the modeled topics is based on certain probabilities and helps towards assigning each document with the identified topics.

Non-negative Matrix Factorization (NMF) NMF is a linear-algebra optimization algorithm used for dimensionality reduction and data analysis [18]. NMF factorizes a document-term matrix (i.e., a matrix representing the frequency of terms in different documents) into two matrices namely the term-feature and the feature-document, with the property that all the three matrices will have non-negative elements [19].

2.3 Relevance Ranking

A paper is considered to be most influential if it is cited more often by other influential papers in the scholarly network [20]. Papers are ranked similar to search engines ranking of web pages, with the difference that instead of using the hyperlink network, the citation network formed by the publications is utilized. Consequently, this ranking mechanism also helps to determine the most important authors. In SNA, determining the most influential authors based on a ranking mechanism on a collaboration network is also possible. Ranking mechanisms based on the page rank algorithm [21], considering collaboration and citation networks were utilized for our analysis. We used 20 iterations, with a damping factor of 0.85.

2.4 Self-Citation Detection

A self-citation occurs when a cited publication shares at least one common author with the publication that cites it [22]. Self-citations boost a paper's citation count for a paper. Self-citations that are introduced in a new research paper to indicate an *extension* to the author's previous work is considered valid. However, unveiling these *semantic self-citations* would require exhaustive processing. In SNA, understanding these self-citation counts for a paper uncovers information on whether the authors have attempted to have a global information coverage on the topic. Considering this aspect, in our work we study if self-citations have an impact on the relevance and citation counts of the most relevant papers.

3. SNA ON CODD'S WORLD

In this section, we highlight important aspects of the SNA performed on Codd's World. We start with the illustration of how Codd's World was created, further describe the formulated research questions, highlight the tooling framework and then present the detailed output evaluation.

3.1 Creation of Codd's World

We choose the bibliographic database for computer science DBLP [8] as the main source of information with the related *Paper Abstracts* curated from the Microsoft Academic Graph [23]. We used the DBLP release of April 2018. To narrow down information in DBLP specific to the database community, we take the node of the foundational paper of Edgar Codd on relational databases [10]. Furthermore, all papers with transitive reference relationships to the main paper node were considered (i.e., papers that cite transitively the work of Codd), with no limitations on path length. This leads to the formation of our database community citation network graph (Codd's World). Based on this, further sub-graphs were constructed to form the base of the scholarly network. The resulting heterogeneous network consisted of four nodes, namely, (a) authors, (b) papers, (c) venues and (d) journals (cf. Figure 2). The five types of relationships in the network consisted of authorship, collaboration, belonging to venue, belonging to journal and citation. Taken together, these nodes and relationships formed the database of our scholarly network[1], which was then further used

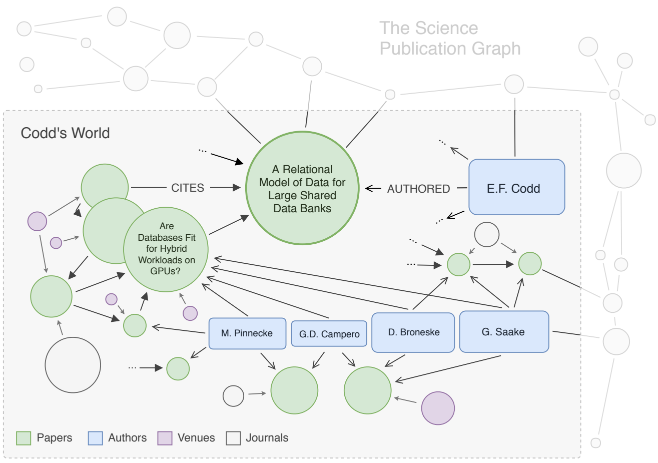

Figure 2: The heterogeneous network *Codd's World*, a subset of the science publication graph: papers (transitively) citing the relational database foundation paper by E.F. Codd, authors of these papers, related venues and journals.

for our SNA. Overall, our scholarly network contains 3,122,404 papers, 1,338,357 authors, 25,166,959 citations and 2,815,781 co-authorship edges.

3.2 Research Questions

Through our SNA, we answer five carefully formulated research questions to understand the scholarly network data. The research questions answered throughout our analysis of the network data are:

- **RQ$_1$**: From topic evolution through time, are there stand-out topics?
- **RQ$_2$**: Does the use of self-citations have an impact on the most cited papers per topic per year?
- **RQ$_3$**: Does the use of self-citations have an impact on the most influential papers per topic per year?
- **RQ$_4$**: Does the use of self-citations have an impact on the citations per topic per year?
- **RQ$_5$**: Is their a difference among the most important authors per topic, looking at collaboration only, citation only, and mixed, while considering self-citations or not?

We further describe the utilized tooling and analysis framework and then present a detailed evaluation of the research questions.

3.3 Tools

We now describe the tools utilized towards answering the formulated research questions through the description of the topic modeling framework with its various aspects concerning topic model selection and the optimal number of topics estimation. Additionally, further analysis carried out is also detailed.

Topic Modeling Framework Since the main aim of the analysis was the extraction of meaningful and frequent topics from the paper abstract, a framework for Topic Modeling was implemented in Python 3.0[2]. The framework utilizes the Python library TOM (TOpic Modeling) [24] for topic extraction using LDA or NMF. TOM library was selected after a careful examination of its offered functionality and the relevant research study utilizing it [25]. This developed framework was mainly used to visualize and answer RQ$_1$.

Topic Model Selection As presented in Section 2.2, there are two approaches that can be used for topic modeling. We applied both techniques, NMF and LDA, and inspected the output of both the TMs on our data set. We found out that the topics given by NMF provided better understanding and were more interpretable when compared to the topics given by LDA. Hence, we use NMF throughout this work for topic modeling.

Optimal Number of Topics Estimation Estimating the optimal number of topics is a crucial part of topic modeling. If the number is too small the topics will be vague and if the number is too large the topics will be redundant and overlap too much. The estimation of the ideal number of topics as 30 was based on the size of the input dataset, following the Greene metric over top 10 words.

Further Analysis The further analysis on the dataset to answer RQ$_2$ - RQ$_4$, used the graph database Neo4j[3]to store and retrieve connected data using its query language, Cypher. The nodes and their relationships were loaded into Neo4j and queries written in Cypher were utilized to obtain results for the respective research questions. For the Page Rank Algorithm required for RQ$_3$ and RQ$_5$, to identify the most influential paper and most influential author respectively, we employed the default provided by Neo4j. The scores were calculated with a Cypher query which implements the page rank algorithm in Neo4j. In order to find the most influential paper and most influential author, the graph of Codd's World was enriched with the calculated page rank scores and fed back into Neo4j.

Design and Assumptions The following assumptions were made prior to the analysis:

(i) The input dataset used for analysis only consists of papers that cite transitively the work of Codd and is assumed to have no duplicate authors.

(ii) The tools and techniques used for analysis are assumed to serve their purpose efficiently. Accordingly, the analysis results were formulated and validated based on the best of the knowledge of the authors.

Considering these design and assumptions we further analyzed the scholarly network.

3.4 Evaluation

In this part, we analyze the results of our SNA in order to answer our research questions RQ$_1$ - RQ$_5$. However, due to space limitations, we have included the detailed output results of the SNA along with the related Cypher queries as a separate technical analysis report which is available online[4].

3.4.1 *RQ$_1$: Evolution of Topics Through Time*

Relevance Visualizing topic evolution depicts popular research topics, and measures topic change over time, increase or decrease of importance for a topic and other topic evolutionary characteristics, thus helping to better understand the research trend in the

[1] Datasets available online:
ftp://itidbftppublic:dbse0Run41rP@ftpiti.cs.uni-magdeburg.de/upload/protolabs/datasets/as-is/codds-world/2019-02

[2] Code available online:
https://github.com/Rspawar/Scholarly-Network-Analysis.git

[3] https://neo4j.com/developer/

[4] https://github.com/Rspawar/Scholarly-Network-Analysis/blob/master/Technical_Analysis_Report.pdf

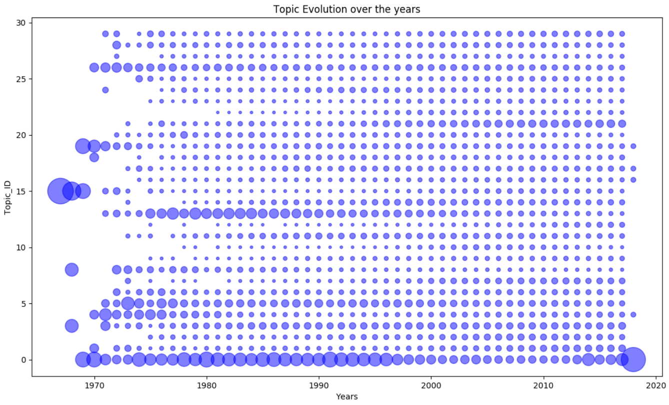

Figure 3: Evolution of topics through time

database field.

Results and Discussion Figure 3 shows the evolution of the identified 30 topics over the years, indicating the percentage of papers from those published in a year that are assigned to a given topic. It is seen from the figure that only on starting years and end years, does there seem to be a disproportion towards some topics. This is due to the data collection and can be dismissed. Some topics, specially Topics 0 (named *Numerical Analysis*) (concerning estimations and cost models) and Topics 5, 11, 19, 26, 28 (*Hardware, Algorithms, Temporal Analysis, Distributed Systems, Information Retrieval*) have seen a steady presence in research throughout the years. Other topics, like 2, 14, 18, 21, 22 (*Networking, Image Processing, Network Analysis, Machine Learning, Video Processing*) have seen an increase in relevance, with Machine Learning's increase being remarkable. Among the topics that have seen a relative decrease interest is Topic 8 (*Operating Systems*), which played a larger role in the community during the first decades shown, but seems to do less so in recent years.

3.4.2 *RQ$_2$: Top Citations per Topic per Year*

Relevance Understanding most cited papers helps to measure the overall scientific impact made by a paper. Recognizing the most cited papers per topic per year facilitates deep analysis through measuring the trends in the scientific impact along the years.

Results and Discussion Cypher queries were executed to return the most cited papers for a particular year with and without self-citation. Tables 1 and 2 summarize the query output for the year 1970 with and without self-citation. Similarly, the network was queried for the year 2017. Comparing the tables, it is observed that the majority of the returned papers with their topics are the same in both the queries. This suggests that the top papers returned do not achieve their most cited criteria through self-citation.

Title	TopicName	Count
A Survey of Analytical Time-Sharing Models	NumericalAnalysis	3
A relational model of data for large shared data banks	DataMining	3
Optimizing the Performance of a Drum-Like Storage	TemporalAnalysis	2
Principles of Optimal Page Replacement	Optimization	1

Table 1: Most cited papers in 1970 with self-citation

Title	TopicName	Count
A Survey of Analytical Time-Sharing Models	NumericalAnalysis	3
A relational model of data for large shared data banks	DataMining	3
Optimizing the Performance of a Drum-Like Storage	TemporalAnalysis	2

Table 2: Most cited papers in 1970 without self-citation

3.4.3 *RQ$_3$: Top Influence per Topic per Year*

Relevance Measuring the most influential paper based on its ranking in the network is an indicator of high acceptance of the research work by the scientific community. Visualizing the top influential papers per topic helps to understand the trend of topic acceptance over the years.

Results and Discussion Cypher queries were executed to return the most influential papers (based on Page Rank score) for a particular year with and without self-citation. Tables 3 and 4 summarize the query output for a particular year 1970 with and without self-citation. Similarly, the network was queried for all the years and also on specific years like 2017. Observation of the tables suggests that the highest Page Rank is indeed associated with the old papers but is not necessarily always true. As expected, the foundational paper of Edgar Codd on relational databases remains the most influential over all the years (with and without self-citation). The results of this research question cannot be compared with the results of RQ$_2$ as self-citation makes a difference on the network dynamics (given that Page Rank scores depend on the complete network structure) but not on the citation count of the most cited papers. Furthermore, we observe that removing self-citations leads to a higher range for the scores of the most influential paper, showing that self-citation does indeed make a difference in the scoring, though it does not change the top items ranked in their influence.

Title	Topic Name	Score
A relational model of data for large shared data banks	DataMining	814.42
Virtual memory	NumericalAnalysis	151.17
Toward an understanding of data structures	NetworkAnalysis	27.37
A schema for describing a relational data base	NumericalAnalysis	18.15
Introduction to storage structure definition	NumericalAnalysis	3.31
Time-sharing for OS	TemporalAnalysis	1.64
TICKETRON: a successfully operating system without an operating system	DistributedSystems	0.23
Swap-Time Considerations in Time-Shared Systems	TemporalAnalysis	0.18
A contiuum of time-sharing scheduling algorithms	Applications	0.15

Table 3: Most influential papers (based on Page Rank score) with self-citation 1970

3.4.4 *RQ$_4$: Citations per Topic Through Time*

Relevance Measuring citation count for a topic helps to understand its research popularity among the scientific community. Analyzing citation count per topic per year helps to measure the relevant trends of research on a topic over the years.

Results and Discussion Cypher queries were executed to return the citation count for a topic for all the years. Figure 4 visualizes the output for topic Machine Learning. Observing Figure 4

Title	Topic Name	Score
A relational model of data for large shared data banks	DataMining	13669.49
Toward an understanding of data structures	NetworkAnalysis	5092.02
Virtual memory	NumericalAnalysis	3988.57
A schema for describing a relational data base	NumericalAnalysis	264.00
Introduction to storage structure definition	NumericalAnalysis	18.38
TICKETRON: a successfully operating system without an operating system	DistributedSystems	12.22
Time-sharing for OS	TemporalAnalysis	5.36
Swap-Time Considerations in Time-Shared Systems	TemporalAnalysis	0.21
A contiuum of time-sharing scheduling algorithms	Applications	0.15

Table 4: Most influential papers (based on Page Rank score) without self-citation 1970

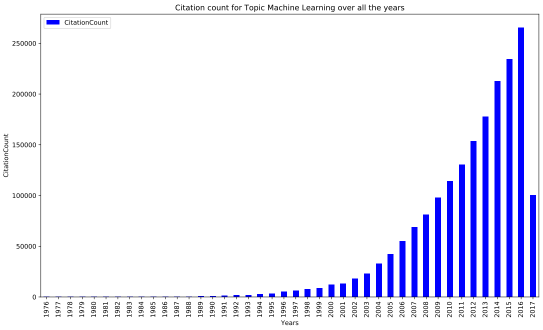

Figure 4: Citation count for topic Machine Learning over all the years

indicates an increasing trend for the selected topic Machine Learning over the years. Additionally, running the query for other topics identified no significant downtrend for any topic. This could be given the fact that we have not included information regarding the distribution of the papers having high citation counts. Power Law analysis [26] can be used to solve this problem in the future by drilling down into papers.

3.4.5 RQ_5: Top Influential Author per Topic

Relevance Measuring the top influential author per topic combined and ranked over all the years, helps to understand the popular acceptance of the author's research on a particular topic among the scientific community. It is also an indicator of the valuable contribution made by the author towards the research topic.

Results and Discussion Cypher queries were executed to return the most influential authors through combination of Author Rank and Page Rank on collaboration/co-authorship network for a particular topic with and without self-citation. We found that Author Rank alone did not provide insightful information on the top influential authors, with possible errors introduced by lack of author name disambiguation. We also found little differences when combining Page Rank and Author Rank, with respect to Page Rank alone, since the scores on the collaboration network are in a smaller scale when compared to the scores on the citation network. Table 5 and 6 summarize the partial output of the top authors according to a combined AuthorRank and PageRank score, for the topic Data Mining with and without self-citation. Results show that removing self-citation has a large influence on the combined score, affecting the order of top items in the list. This shows that, though these kind of citations do not affect the ranking of top papers, the sum of these changes as it is aggregated into the score of an author, creates a difference. Hence self-citation does influence the ranking of the top authors, therefore scoring measures that dismiss these kind of links, could be a good choice for understanding authors relevance in the network we study.

Author Name	Score
Scott Shenker	2323.37
Demetri Terzopoulos	1693.30
Geoffrey E. Hinton	1563.39
Hari Balakrishnan	1534.55
Rakesh Agrawal	1505.22

Table 5: Combination of Author Rank and Page Rank with self-citation for Topic Data Mining

Author Name	Score
E. F. Codd	18399.13
Daniel G. Bobrow	14275.87
Carl Hewitt	12347.67
Ben Wegbreit	9271.63
Peter J. Denning	7198.57

Table 6: Combination of Author Rank and Page Rank without self-citation for Topic Data Mining

4. RELATED WORK

SNA towards an understanding of the scientific research community can be seen as a fast growing research area. This analytical study carried out and presented in our paper is mainly inspired by Prof. Dr. Erhard Rahm and Prof. Dr. Andreas Thor work on Citation analysis of database publications [9] and Prof. Dr. Erhard Rahm and David Aumüller work on Affiliation analysis of database publications [27]. Additionally, Topic Modeling for SNA helps to extract meaning out of the scholarly information in the form of research topics. The work which is closely related to our research [7] is based on a hybrid topic model. Contrary to our research, where it is required to set a fixed number of topics, the hybrid topic model incorporates a dynamic model which is not based on a fixed number of topics. This dynamic consideration can further prove beneficial to accurately model and understand the evolving and ever-changing nature of the scientific community.

5. CONCLUSION AND FUTURE WORK

Summarizing, in this paper we introduce the Codd's World dataset, and we presenting some early analytical results obtained through SNA on this dataset. The modeling of the 30 topics presented the prominent areas of research in the database community. Interestingly it is observed that the database community, when considered as an influence network in Codd's World, includes many topics that go beyond databases, such as Machine Learning, Energy and a few others. This shows the large influence of database research on overall computing research. From our evaluation we can also report a relative downtrend in publications for some topics like Operating Systems, and uptrends in Machine Learning and Image Processing. Similarly, we report that no topic has a dominance in their presence throughout the years, though some topics like NumericalAnalysis (concerning cost models in databases) and Hardware show a constant presence through time. Similarly we show that self-citation plays no big role on the most cited papers, though it affects the network structure and thus changes the scoring, having impact on the long tail of cited papers. We also find results suggesting that to date, the collaboration network that we study is not sufficient to evaluate an authors relevance, but that instead a ranking based on citation networks while removing self-citations could be a good choice.

Future work could improve limitations in this dataset, observed from authors sharing the same name but being different, which might have introduced errors in the ranking of authors in the collaboration network. Future work can also consider other dimensions for SNA on this dataset.

6. ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their insightful feedback. This work was partially funded by the DFG (grant: SA 465/50-1).

7. REFERENCES

[1] Feng Xia, Wei Wang, Teshome Megersa Bekele, and Huan Liu. Big scholarly data: A survey. *IEEE Transactions on Big Data*, 3(1):18–35, 2017.

[2] Erjia Yan and Ying Ding. Scholarly networks analysis. In *Encyclopedia of Social Network Analysis and Mining*, pages 1643–1651. Springer, 2014.

[3] Mark Newman. *Networks*. Oxford university press, 2018.

[4] Wei Wang, Jiaying Liu, Shuo Yu, Chenxin Zhang, Zhenzhen Xu, and Feng Xia. Mining advisor-advisee relationships in scholarly big data: A deep learning approach. In *Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries*, pages 209–210. ACM, 2016.

[5] Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. Topic modelling for qualitative studies. *Journal of Information Science*, 43(1):88–102, 2017.

[6] Abram Hindle, Michael W Godfrey, and Richard C Holt. What's hot and what's not: Windowed developer topic analysis. In *Software Maintenance, 2009. ICSM 2009. IEEE International Conference on*, pages 339–348. IEEE, 2009.

[7] Christin Katharina Kreutz. A hybrid approach for dynamic topic models with fluctuating number of topics. In *Grundlagen von Datenbanken*, pages 35–40, 2018.

[8] Michael Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *International symposium on string processing and information retrieval*, pages 1–10. Springer, 2002.

[9] Erhard Rahm and Andreas Thor. Citation analysis of database publications. *ACM Sigmod Record*, 34(4):48–53, 2005.

[10] Edgar F Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[11] Satrio Baskoro Yudhoatmojo and Muhammad Arvin Samuar. Community detection on citation network of dblp data sample set using linkrank algorithm. *Procedia Computer Science*, 124:29–37, 2017.

[12] Kevin W Boyack and Richard Klavans. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12):2389–2404, 2010.

[13] Tze-Haw Huang and Mao Lin Huang. Analysis and visualization of co-authorship networks for understanding academic collaboration and knowledge domain of individual researchers. In *Computer Graphics, Imaging and Visualisation, 2006 International Conference on*, pages 18–23. IEEE, 2006.

[14] Qin He. Knowledge discovery through co-word analysis. 1999.

[15] Erjia Yan and Ying Ding. Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, 63(7):1313–1326, 2012.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[17] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[18] Michael W Berry and Murray Browne. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11(3):249–264, 2005.

[19] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[20] Carl Bergstrom. Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, 68(5):314–316, 2007.

[21] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[22] Yuxiao Dong, Hao Ma, Zhihong Shen, and Kuansan Wang. A century of science: Globalization of scientific collaborations, citations, and innovations. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1437–1446. ACM, 2017.

[23] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM, 2015.

[24] Adrien Guille and Edmundo-Pavel Soriano-Morales. Tom: A library for topic modeling and browsing. In *EGC*, pages 451–456, 2016.

[25] Adrien Guille, Edmundo-Pavel Soriano-Morales, and Ciprian-Octavian Truica. Topic modeling and hypergraph mining to analyze the egc conference history. In *EGC*, pages 383–394, 2016.

[26] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[27] David Aumüller and Erhard Rahm. Affiliation analysis of database publications. *ACM SIGMOD Record*, 40(1):26–31, 2011.