

Comparison Analysis of Co-authorship Network and Citation Based Network for Author Research Similarity Exploration*

Jeeyoung Yoon (윤지영)**

Min Song (송민)***

Contents

- | | |
|-----------------|---------------|
| 1. Introduction | 4. Result |
| 2. Related Work | 5. Conclusion |
| 3. Methodology | |

ABSTRACT

Exploring research similarity of researchers offers insight on research communities and potential interactions among scholars. While co-authorship is a popular measure for studying research similarity of researchers, it cannot provide insight on authors who have not collaborated yet. In this work, we present novel approach to capture research similarity of authors using citation information. Extensive study is conducted on DATA & KNOWLEDGE ENGINEERING (DKE) publications to demonstrate and compare suggested approach with co-authorship based approach. Analysis result shows that proposed approach distinguishes author relationships that is not shown in co-authorship network.

Keywords: Author research similarity analysis, Co-authorship network, Citation based author analysis, Network Analysis, Data & Knowledge Engineering

* This work was supported by a National Research Foundation of Korea grant funded by the Korean government (NRF-2020S1A5B1104865)

** Graduate Student at Department of Library and Information Science, Yonsei University (jeeyoungyoon9@yonsei.ac.kr) (First Author)

*** Professor at Department of Library and Information Science, Yonsei University (min.song@yonsei.ac.kr / ISNI 0000 0000 8203 196X) (Corresponding Author)

논문접수일자: 2022년 10월 18일 최초심사일자: 2022년 11월 4일 게재확정일자: 2022년 11월 25일
한국문헌정보학회지, 56(4): 269-284, 2022. <http://dx.doi.org/10.4275/KSLIS.2022.56.4.269>

※ Copyright © 2022 Korean Society for Library and Information Science

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. Introduction

Studying research similarity of authors are significant for understanding the research landscape of a domain. The author research similarity can help discover research communities and predict the interactions between authors (Luong et al., 2015). With such importance, different attempts have been made to investigate the research similarity of authors (Yan & Ding, 2009; Liu et al., 2005; Lima et al., 2020). One of the most widely used methods for such inquiry is co-authorship analysis. Co-authorship of authors occur when authors jointly appear in a paper, and we can construct co-authorship networks based on these types of links between two scientists (Uddin et al., 2012). While co-authorship is a robust measure for capturing the research similarity of authors, it is limited in that it cannot distinguish author pairs who have not collaborated yet. In the present study, we present novel approach to distinguish author research similarity based on citation information. To be specific, presented approach captures research similarity of authors based on frequency of being cited together by same work. To compare co-authorship based approach and presented citation based approach, we perform extensive analysis on DATA & KNOWLEDGE ENGINEERING (DKE) papers. DKE is a monthly peer-reviewed academic journal which is well known to worldwide scholarly communities. Ideas on database systems and knowledgebase systems are being actively exchanged via DKE. To conduct the analysis, we first construct co-authorship network and distinguish author communities. Keywords presented by authors are used to understand the research interest of the communities. Then, we conduct citation based author research similarity analysis and compare the result with the co-authorship counterpart.

The rest of this paper is constructed as follows. First, we summarize past works on co-authorship analysis and citation based author analysis. Then, we explain our research method, followed by result section. We conclude our work with conclusion section.

2. Related Work

This section reviews previous literatures focusing on co-authorship networks and citation based author analysis.

2.1 Co-authorship network

The authors who have collaborated on the same paper may have a higher research similarity

than others (Zhang et al., 2018). Moreover, the numeric features of scientific collaboration can be reliably tracked via this approach (Glänzel & Schubert, 2004). Two authors are considered to have a co-authorship relationship if they appear jointly in a paper, and co-authorship networks are formed based on these types of links between two scientists (Uddin et al., 2012). Thus, networks of scientists are constructed, where edges are formed if two scientists (nodes) co-authored a paper (Newman, 2004). The analysis of co-authorship network aims to investigate the phenomenon of scientific collaboration between authors.

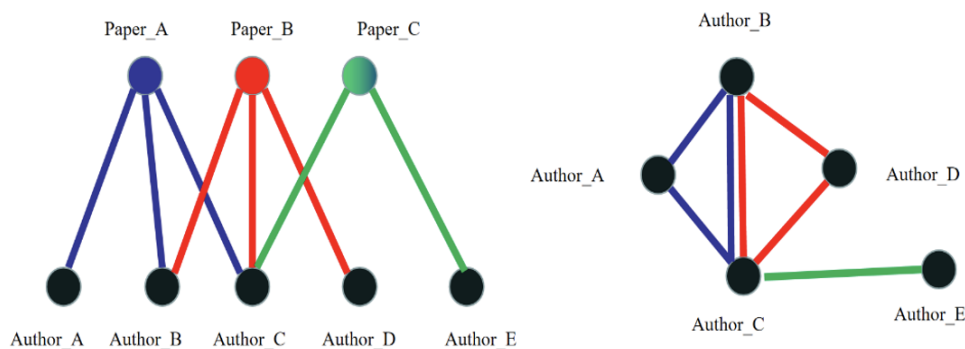
Research on co-authorship networks has been actively conducted. Yan & Ding (2009) calculated four centrality measures (closeness centrality, betweenness centrality, degree centrality, and PageRank) for authors in the constructed co-authorship network, and they found that centrality measures can be valuable indicators for impact analysis. Liu et al. (2005) presented author rank, which is an indicator of influence of individual authors in the network. They found that author rank can be an alternative metric to evaluate research influence. Lima et al. (2020) also performed an in depth analysis, where they discovered that the degree centrality metric could help pinpoint highly prolific authors who have been active members of the community.

The constructed network can be considered as an interdisciplinary scientific collaboration network because the authors are from different disciplines, and their research communities exist separately. From this perspective, Liu & Xia (2015) unraveled the structure and evolution of the co-authorship network in an interdisciplinary research field. Feng & Kirkley (2020) also assessed collaborative preferences in interdisciplinary research at multiple scales by examining disciplinary mixing patterns. Their results demonstrate that disciplinary diversity is reflected by the diverse research experiences of individual researchers rather than diversity of pairs or groups of researchers. Huang & Chang (2011) investigated interdisciplinary changes in information sciences (from 1978 to 2007). They found that the co-authors of information science articles are primarily from the discipline of library and information science. Nam & Park (2014) studied the factors influencing the research collaboration between scholars using co-author networks. They presented that geographical proximity and research similarity between authors significantly influence the collaboration dynamics. Co-authorship network can also be used to construct network between institutions. Kim et al. (2021) expanded the idea of co-authorship network to study research collaboration in an institutional level.

As such, collaboration of research community can be studied by constructing a co-authorship network. Hence, communities can be found using network algorithms, and the core author pairs with different research similarities in each community can be discovered by co-authorship network analysis.

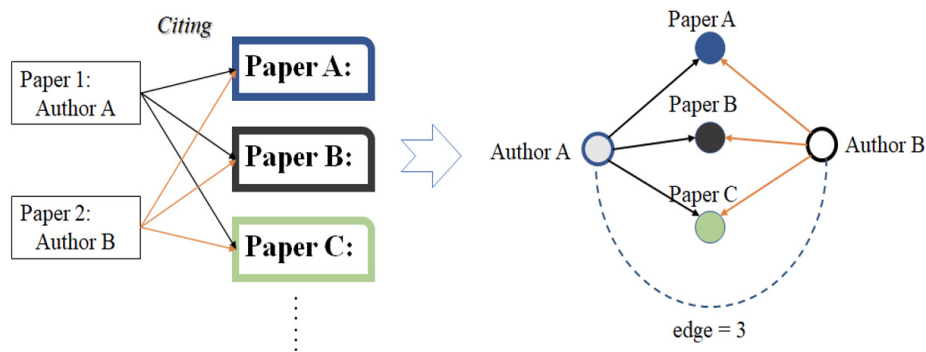
2.2 Citation Based Author Analysis

Another valuable resource we can gain insight on authors is citation information. There have been attempts to analyze author relationship based on reference list provided by papers. Citation based author networks can be divided into two: author co-citation networks and bibliographic coupling (BC) networks. In author co-citation networks, the authors' publication generally has a time lag to appear in citation based author co-citation studies; thus, it is hard to detect currently active authors via author co-citation network. A BC network can address this problem, as the cited reference list appears when the paper is published. The BC method was first introduced by Kessler (1963) to compute the relationship between papers that cite the same paper, and the strength weight of BC between two papers is represented by the number of documents shared in their reference lists. In <Figure 1>, Papers A, B, and C are the cited papers in the reference list, and the authors of the citing papers are Authors A, B, C, D, and E. The network projection is as follows:



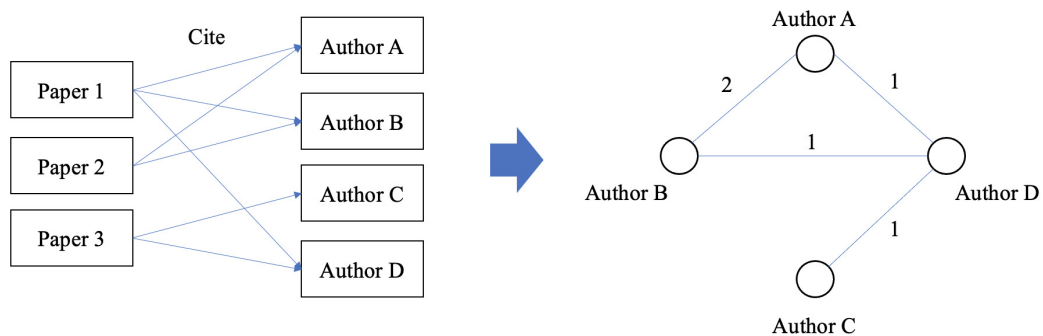
<Figure 1> Author-bibliographic coupling bipartite network projection

The ABC network (Zhao & Strotmann, 2008) represents the relationship between two authors who have cited same articles, extending the concept of BC from paper level to author level. Based on the assumption of ABC, the more references two authors have in common, the more similar their research is. The coupling strength is the number of cited articles that two citing authors share. The ABC network displays the relations of authors, as BC indicates relatedness between papers (Zhao & Strotmann, 2014). If two authors are cited from same papers, indicating higher BC counts, they may have higher research relatedness. <Figure 2> presents the author-bibliographic coupling network.



〈Figure 2〉 Author-bibliographic coupling network

Author research similarity calculation of this work is inspired by this concept. In this work, if authors are cited by same paper, edge between authors gain weight. 〈Figure 3〉 presents the suggested method. The edge between nodes indicates the number of co-citing papers. We collected all reference information for each paper in the dataset and compared them to scrutinize the number of papers that cites the same author pairs. Based on such information, undirected network of author similarity is constructed.



〈Figure 3〉 Citation based author similarity network (proposed approach)

3. Methodology

The primary aim of this paper is to compare the author similarity based on co-authorship and citation. To do so, author community analysis via co-authorship network construction and citation based author research similarity analysis is conducted. Suggested method is demonstrated using

the DKE dataset.

3.1 Data Collection

We use the Scopus database¹⁾ to collect the DKE dataset of full-text papers in PDF file format. Searching within the source title “*Data & Knowledge Engineering*,” the results reveal published documents for 36 years (1985 to 2020). Total of 1,650 full-text papers (1,527 article papers and 123 conference papers) are selected after filtering editorial, erratum, conference reviews, and reviews. The number of papers for each year is presented in <Figure 4>.



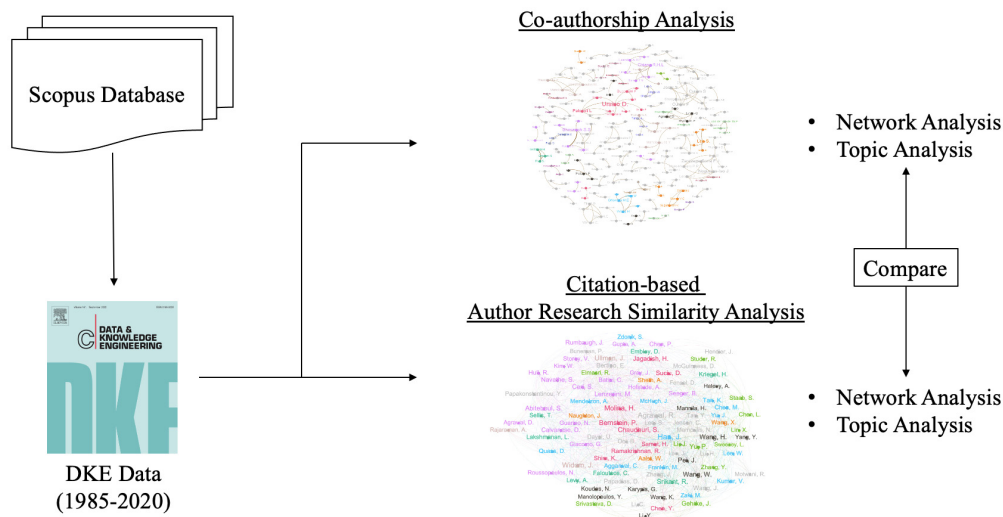
<Figure 4> Number of yearly DATA & KNOWLEDGE ENGINEERING (DKE) published papers

We also gathered the bibliographic information. The dataset collection task was accomplished in February 2021. Collected data is used for further analysis, including co-authorship network construction and citation based author research similarity analysis.

3.2 Procedure

We present the overview of our research in <Figure 5>.

1) <https://www.scopus.com>.



〈Figure 5〉 Overview of our research

The full-text data and bibliography data are gathered from the Scopus database. The bibliography data is used to construct the co-authorship network. With the constructed network, we conduct author community analysis to understand the network dynamics and research interest of formed communities. In addition, reference data are utilized to reveal the author research similarity based on citation. To be specific, we use citation information of papers to distinguish author pairs which are cited in same work with high frequency. Result of co-authorship network analysis and citation based author similarity analysis is compared to see whether suggested approach captures author relationships which are not distinguished by the former.

3.3 Co-authorship Network Analysis

Analyzing co-authorship relationship is a popular approach for studying research interest of scholars. Therefore, we construct co-authorship network to observe the relationship between authors and distinguish author communities. To do so, bibliometric data is processed by Python to construct a GraphML file. This GraphML file is then visualized and analyzed using Gephi (Bastian, Heymann, & Jacomy, 2009). First, we visualize the network to understand the general landscape of co-authorship dynamics. Then, we conduct modularity analysis to distinguish author communities. With modularity analysis, we can determine subclasses which construct the network (Khokhar, 2015). In this work, Louvain algorithm (Blondel et al., 2008) is used for modularity analysis. After distinguishing

author communities, research interest of each subclass is studied by research key words provided by authors.

3.4 Citation Based Research Similarity Analysis

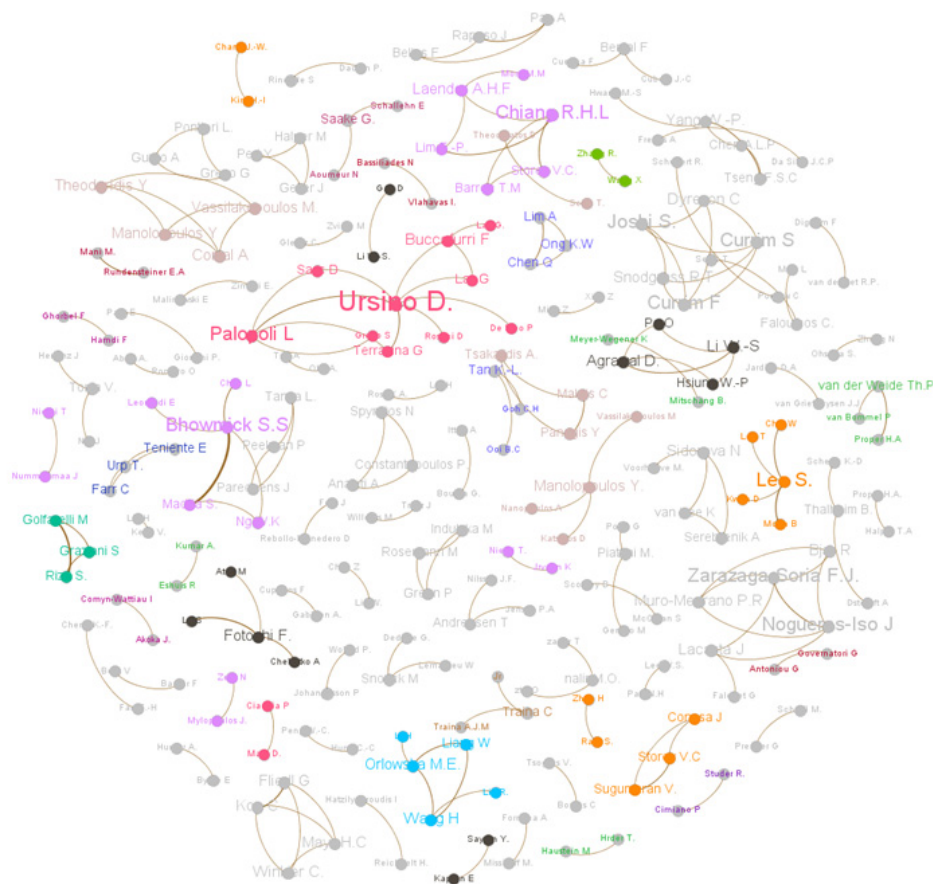
To compare research similarity captured by co-authorship with that embedded in the citation information of papers, this work conducts citation based author research similarity analysis. This approach is inspired by the idea of ABC network, where similarities of authors are scrutinized by the works they cite. In this work, similarities between authors are weighted based on the number of papers that cite the author pairs together. To do so, reference list is used to construct a list of authors cited by the given paper. Using this list, we accumulate similarity weight. To be specific, authors gain weight in each instance when they are cited by same paper. Authors who are cited together by larger number of papers will gain high research similarity using this approach. Research similarity based author network is then constructed as a GraphML file. Described processes are conducted using Python. After preparing the GraphML file, network analysis using Gephi is performed. Also, to check the research interest similarity of paired authors, research interest of authors provided by Google Scholar is utilized.

4. Result

4.1 Co-authorship network analysis

We attempt to identify authors' communities with similar research interest through co-authorship network analysis. The authors from each paper are extracted to create co-authorship network. In this network, nodes represent authors, while edges represent the co-authorship relation between two authors. We construct the network using the DKE bibliography data, and we find total of 3,519 nodes and 5,602 edges in the co-authorship network. By using Gephi, the whole network data is visualized to detect the author communities in DKE papers.

<Figure 6> shows a DKE co-authorship network visualization graph that consists of nodes (authors) and edges (co-authorship relations where the edge degree is greater than 2). The color represents the author communities. The fact that authors have higher relatedness in co-authorship network may indicate they have higher research similarity (Zhao & Strotmann, 2008).

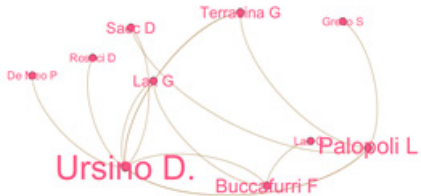
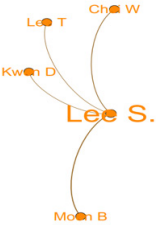




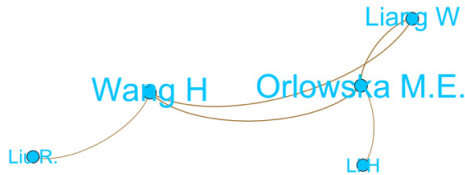
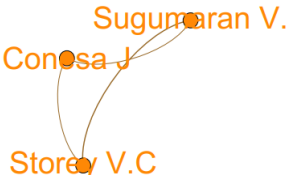
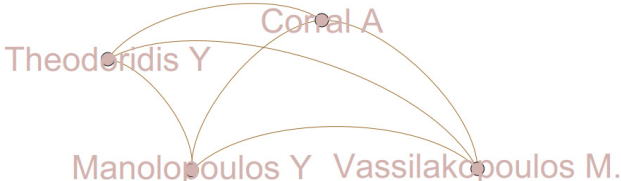
<Figure 6> The co-authorship network visualization (edge degree > 2)

Keywords given by the authors can be seen as the concentrative summary of the publication. Keywords provide understandable content to represent author's research topic (Li et al., 2015). Thus, we group authors that appeared in the same author community and collect the keywords of each author. For each collaboration community, we list the top ten keywords ranked by frequency to represent the authors' research topics. <Table 1> summarizes the research topics of seven main collaboration communities which was shown in <Figure 6>. For the collaboration community which is constituted by authors including Ursino, D., Buccafurr, F., Palopoli, L. and Lax, G., the key research topic is database scheme integration, semantic, conceptual data models, and ontologies. Lee S is the main author for second community, where the research topic includes moving objects, spatio-temporal databases, and index structures. The largest node in third community is the author Chiang, R.H.L.. The community in general is interested in database design, database

reverse engineering, and entity-relationship model. Three authors are listed in the fourth community: Graziani, S., Rizzi, S. and Golfarelli, M.. Research topics of this community includes data warehousing, olap, view materialization, and cross-version querying. Wang, H. and Orłowska, M.E. are the main authors of the fifth community. Research topics of this community includes data integration, design of data warehouse, heuristic algorithms, and security. Author Conesa, J., Sugumaran, V. and Storey, V.C. are the key authors of the sixth community. Their research topics include ontology, database design and entity-relationship model, which are similar to that of the third community. The seventh collaboration community includes authors Manolopoulos, Y., Vassilakopoulos, M., Corral, A. and Theodoridis, Y.. This community is interested in r-tree, performance analysis, spatial databases, branch-and-bound algorithms, distance join and query processing.

〈Table 1〉 The author research topics of each collaboration communities

<p>Research topics: database scheme integration, hyponymy and overlapping extraction, semi-automatic and semantic methodologies, conceptual data models, knowledge representation techniques, ontologies, semantic heterogeneity, web-based information systems, cooperative work, engineering application</p> 	<p>Research topics: moving objects, spatio-temporal databases, index structures, cell-based access structure, multiversion access structure, overlapping technique, aggregation query, hilbert curve, prefix-sum, spatio-temporal data warehouses, st-cube</p> 
<p>Research topics: database design, database reverse engineering, entity-relationship model, ontology, conceptual modeling, relational databases extended, database design schema, design and evaluation framework, extended entity-relationship model, relational database</p> 	<p>Research topics: data warehousing, olap, view materialization, cross-version querying, schema augmentation, schema versioning, approximate query answering, hierarchical clustering, olam, incremental loading, on-demand etl</p> 

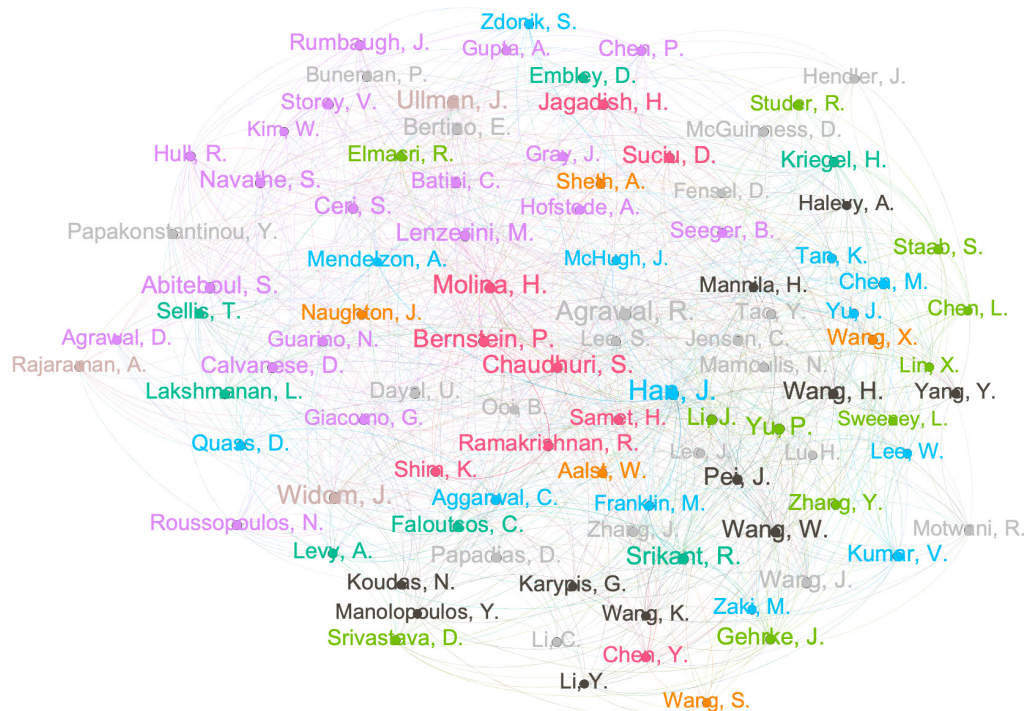
<p>Research topics: data integration, design of data warehouse, heuristic algorithms, incremental maintenance, maintenance time constraint, materialized view selection, security, object dependency, object-oriented databases, propagation policy</p> 	<p>Research topics: ontology, cyc, researchcyc, database design, entity-relationship model, conceptual modeling, ontology auditor, quality metric, semiotic theory, knowledge repositories, query expansion</p> 
<p>Research topics: r-tree, performance analysis, spatial databases, branch-and-bound algorithms, distance join, i/o and response time performance, query processing, buffer model, cost models, distance join queries</p> 	

Co-authorship network analysis shows that collaboration between authors occur in a rather segregated form. The nodes are connected to small number of nodes and intertwined connectivity between nodes is not observed. It is also shown that distinct communities can be formed although research topics of authors are similar (community three and community six).

4.2 Author Research Similarity Analysis

In this section, result of citation based author research similarity analysis is presented. To be specific, research similarity network of authors will be discussed and author pairs with highest research similarity will be studied. Author research similarity network is presented in <Figure 7>. Approximately one hundred nodes are filtered based on degree centrality for visibility of graph. Color of nodes reflect modularity class. While co-authorship network discussed in previous section was a sparse network, research interest network shows more active connection between nodes. Average degree of this network is 9.218. This means that each node is connected to approximately nine nodes in average, indicating that an author shows research similarity with such number of other authors. Also, it is shown that connections between nodes belonging to different modularity

class is also active. This contrasts with co-authorship network where node connection was observed only within the community. Comparison of co-authorship network and research interest network tells us that research interest shared by authors are more intertwined than what are observable by co-authorship relationships.



<Figure 7> Citation based author similarity network (edge degree > 43)

Above network informs us the average connectivity of authors. In other words, quantitative aspect—how many authors share research interest—of research similarity is shown. Studying author pairs with high research similarity tells us the quality of such similarities. Top thirty author pairs with highest research similarity are presented in <Table 2>. Authors who have published their work in DKE and are included in the co-authorship network are bold cased, italicized, and underlined.

〈Table 2〉 Top thirty author pairs with highest similarity weight

Rank	Author Pair		Rank	Author Pair	
1	Rajaraman, A.	Gray, J.	16	Date, C.	Matheus, C.
2	Faloutsos, C.	Fayyad, U.	17	Batini, C.	Premierlani, W.
3	Srikant, R.	Agrawal, R.	18	Schwarz, P.	Mohan, C.
4	Weld, D.	<u>Dunham, M.</u>	19	Stonebraker, M.	<u>Lenzerini, M.</u>
5	Raghavan, P.	Omiecinski, E.	20	Kriegel, H.	<u>Aalst, W.</u>
6	<u>Ceri, S.</u>	Eddy, F.	21	Shim, K.	Ramakrishnan, R.
7	Blaha, M.	Freksa, C.	22	Iris, M.	Tonneau, C.
8	Snodgrass, R.	<u>Jarke, M.</u>	23	Ghelli, G.	Robson, D.
9	<u>Franconi, E.</u>	<u>Kim, W.</u>	24	<u>Navathe, S.</u>	<u>Ceri, S.</u>
10	Stonebraker, M.	Premierlani, W.	25	Lakshmanan, L.	Wiener, J.
11	Wiederhold, G.	<u>Mannila, H.</u>	26	Navathe, S.	<u>Storey, V.</u>
12	Carey, M.	<u>Kim, W.</u>	27	Blaha, M.	Garza, J.
13	Bernstein, P.	Ramakrishnan, R.	28	Franklin, M.	<u>Han, J.</u>
14	Han, J.	Ester, M.	29	<u>Naughton, J.</u>	Noy, N.
15	<u>Navathe, S.</u>	<u>Batini, C.</u>	30	<u>Shapiro, G.</u>	<u>Tanca, L.</u>

Comparison of above table and co-authorship information tells us that none of the author pairs in the above table showed actual co-authorship. Rajaraman, A. and Gray, J. showed the highest research similarity. Research interest of each author are data, algorithms, AI and databases respectively. Faloutsos, C. and Fayyad, U. both shows interest towards data mining. Snodgrass, R. and Jarke, M. are both interested in database systems. As such, author pairs constructed by the proposed approach showed tendency of having similar research interest. Also, authors showed high research similarity although they do not actually involve in co-authorship. One may question that this may be a result from lack of preference on publishing at DKE, however, such trend is also shown by authors who are active in DKE co-authorship community. To be specific, authors in pairs (Franconi, E. & Kim, W.), (Navathe, S. & Batini, C.), (Navathe, S. & Ceri, S.), and (Shapiro, G. & Tanca, L.) all conducted scholarly activity in DKE and showed high research interest similarity, however, did not conducted any collaboration. Citation based author similarity analysis result shows that there exists author pairs where high research interest is shared but do not possess co-authorship. We believe that the proposed approach can contribute to understanding the similarity of authors who have not collaborated yet and provide insight on potential collaborations.

5. Conclusion

Studying research similarity of researchers provides insight on current research landscape and potentials for collaboration. Popular measure for studying research similarities between authors is co-authorship. However, authors who share similar research interest but have not collaborated before are hard to distinguish with this approach. In this work, novel approach to deduce research similarity based on citation is presented. Proposed approach captures research similarities of authors based on papers that cite both authors, leading to high similarity when authors share larger number of citing papers. To demonstrate our approach, extensive analysis on DKE conference papers is conducted.

Analysis result of our work shows that suggested approach successfully captures similarity dynamics that differ from co-authorship network. To be specific, while co-authorship network resulted in sparse network where most authors are severed from each other, network from our approach results in network with higher connectivity. We also observed that suggested approach distinguishes author pairs which are not identified by co-authorship relationship but possess high research similarity.

Although our work presents interesting results on research similarity of authors, our work has limitation in that it is a single journal based study. In future works, we intend to incorporate larger number of journals to deduce a more comprehensive result on research similarity dynamics of different authors.

References

- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1), 361-362.
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
- Feng, S. & Kirkley, A. (2020). Mixing patterns in interdisciplinary co-authorship networks at multiple scales. *Scientific Reports*, 10(1), 1-11.
- Glänzel, W. & Schubert, A. (2004). Analysing scientific networks through co-authorship. In Moed,

- H. F., Glänzel, W., & Schmoch, U. eds. Handbook of Quantitative Science and Technology Research. Springer, Dordrecht, 257-276.
- Huang, M. H. & Chang, Y. W. (2011). A study of interdisciplinarity in information science: using direct citation and co-authorship analysis. *Journal of Information Science*, 37(4), 369-378.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10-25.
- Khokhar, D. (2015). Gephi Cookbook. Packt Publishing Ltd.
- Kim, D., Kim, K., & Zhu, Y. (2021). A bibliometric analysis of the major Korean journals indexed in 2020 Google Scholar metrics. *Korean Society for Information Management*, 38(1), 53-69.
- Li, F., Li, M., Guan, P., Ma, S., & Cui, L. (2015). Mapping publication trends and identifying hot spots of research on Internet health information seeking behavior: a quantitative and co-word biclustering analysis. *Journal of Medical Internet Research*, 17(3), e3326.
- Lima, L. H. C., Laender, A. H., Moro, M. M., & De Oliveira, J. P. M. (2020). An analysis of the collaboration network of the international conference on conceptual modeling at the age of 40. *Data & Knowledge Engineering*, 130, 101866.
- Liu, P. & Xia, H. (2015). Structure and evolution of co-authorship network in an interdisciplinary research field. *Scientometrics*, 103(1), 101-134.
- Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6), 1462-1480.
- Luong, N. T., Nguyen, T. T., Jung, J. J., & Hwang, D. (2015). Discovering co-author relationship in bibliographic data using similarity measures and random walk model. In *Asian Conference on Intelligent Information and Database Systems*. Springer, Cham, 127-136.
- Nam, E. & Park, J. (2014). Factors Influencing Research Collaboration in the Field of Informetrics. *Journal of the Korean Society for Library and Information Science*, 31(4), 201-227.
- Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5200-5205.
- Uddin, S., Hossain, L., Abbasi, A., & Rasmussen, K. (2012). Trend and efficiency analysis of co-authorship network. *Scientometrics*, 90(2), 687-699.
- Yan, E. & Ding, Y. (2009). Applying centrality measures to impact analysis: a coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10), 2107-2118.

- Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2018). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology*, 69(1), 72-86.
- Zhao, D. & Strotmann, A. (2008). Author bibliographic coupling: another approach to citation-based author knowledge network analysis. *Proceedings of the American Society for Information Science and Technology*, 45(1), 1-10.
- Zhao, D. & Strotmann, A. (2014). The knowledge base and research front of information science 2006-2010: an author cocitation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology*, 65(5), 995-1006.