ScholarNodes: Applying Content-based Filtering to Recommend Interdisciplinary Communities within Scholarly Social Networks

Md Asaduzzaman Noor
Montana State University
Bozeman, Montana, USA
mdasaduzzamannoor@montana.edu

Jason A. Clark
Montana State University
Bozeman, Montana, USA
jaclark@montana.edu

John W. Sheppard
Montana State University
Bozeman, Montana, USA
john.sheppard@montana.edu

ABSTRACT

Detecting communities within dynamic academic social networks and connecting these community detection findings to search and retrieval interfaces presents a multifaceted challenge. We explore an information retrieval method that integrates both partition-based and similarity-based network analysis to identify and recommend communities within content-based datasets. Our prototype "ScholarNodes" web interface bridges the gap between community detection algorithms (Louvain, K-means, Spectral clustering) and the BM25 (Best Matching 25) ranking algorithm within a cohesive user interface. From free-text keyword queries, ScholarNodes recommends collaborations, identifies local and external researcher networks, and visualizes an interdisciplinarity graph for individual researchers using the OpenAlex dataset, a global collection of academic papers and authors. Beyond the specific information retrieval use case, we discuss the broader applicability of the methods to generic social network analysis, community detection, and recommender systems. Additionally, we delve into the technical aspects of generating topical terms, community alignment techniques, and interface design considerations for integrating community detection algorithms into a search experience.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Recommender systems, Collaboration recommendations, Cross-domain recommendation, Network visualization, Social network analysis, Interface Design

ACM Reference Format:
Md Asaduzzaman Noor, Jason A. Clark, and John W. Sheppard. 2024. ScholarNodes: Applying Content-based Filtering to Recommend Interdisciplinary Communities within Scholarly Social Networks. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (SIGIR '24).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

SIGIR '24, July 14–18, 2024, Washington, DC

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Community detection within academic social networks remains a complex problem. The evolving and growing scholarship corpus of academic articles and authors adds to this complexity. Partition-based methods which look to divide a network into related communities based on modularity assignments offer one option for network analysis within these types of academic datasets. Similarity-based methods which group nodes based on similarity measures offer another possibility. In our research, we have found utility in detecting and recommending communities using both methods. The challenge has been in connecting these algorithms (Louvain, K-means, or Spectral clustering) to standard search weighting algorithms, like BM25 (Best Matching) or TF-IDF (Term Frequency - Inverse Document Frequency). To this end, we present ScholarNodes, a content-based filtering web interface that allows for visualizing a scholarship graph for researchers and their scholarship networks.

ScholarNodes recommends topics, identifies local and external researcher networks, and visualizes an interdisciplinarity graph for individual researchers using the OpenAlex dataset, a global collection of academic papers and authors. Users are able to ask free-text queries on topics, names, and keywords to discover connections and recommendations for research collaborations, get an understanding of the domain expertise for a particular grouping of researchers, and match their research interests to the scholarship that exists. Our prototype speaks to a particular academic use case. Even so, the methods can be applied to generic social network analysis, community detection, and recommender systems, and we discuss this broader impact to contextualize the work for the information retrieval community. Generation of topical terms for a search index, community alignment techniques, and the interface design requirements for connecting community detection algorithms to a search experience are also discussed.

2 RELATED WORK

Commercial software systems for higher education exist for competitive intelligence tasks and analysis of research productivity. Elsevier Pure is one example of a Research Information Management System (RIMS) These commercial systems have a market and are usually purchased or used within university analytics and data offices as a behind-the-firewall resource. They are also usually given a set number of users and conditions for running diagnostic queries and reports. In this common scenario, faculty, students, and staff aren't given access to learn about the research happening on their campus. Connections remain invisible and partnerships for research are unrealized. Open-source RIMS provide the opportunity to bring data analytics and recommendations to a broader audience and create visibility for research collaboration and mentorship. [4]

There are examples of previous work to develop collaboration recommendation systems based on researchers' publication content. Liang *et al.* [8] proposed a method for offering cross-disciplinary collaboration recommendations. Additionally, Kong *et al.* [6] introduced the Beneficial Collaborator Recommendation model (BCR), considering researchers' dynamic research interests and academic influence (i.e., impact factor). Both methods encoded content data into a vector space to generate recommendations but did not incorporate the social network aspect. In our approach, we integrated social network analysis with publication content, enhancing the interpretability of the recommender's decision-making.

Other studies adopted a hybrid approach, combining both content data and social network analysis. Kong *et al.* [7] used co-author relationships to construct an academic social network and utilized publication contents to reinforce connections between researchers. Zhou *et al.* [18] proposed a multidimensional academic network analysis using multisource scholar data, introducing time-aware edge relationships. Hybrid approaches typically leverage observable relationships to construct the network and then enhance edge weights with additional data. However, this approach may not be suitable for cross-domain research recommendations where observable relationships may not exist. Therefore, we chose to construct the academic network solely based on topic-based similarity.

3 DATASET

For this demonstration, we used information about current researchers at Montana State University (MSU). Based on this, we were able to collect data on 575 researchers with their name, college (e.g., College of Engineering, College of Letters and Science), and department (e.g., Electrical Engineering, Physics, and Ecology) from the university database.

To identify research publications, we utilized OpenAlex [1] [15], an open-source platform that provides a comprehensive interconnected catalog of scholarly papers, authors, institutions, venues, and more. OpenAlex allows filtering publications based on an Institution identifier, and we used the MSU identifier to obtain research articles from 2004 to 2023, where at least one of the authors had an affiliation with the institution. With that, we were able to extract 16, 827 articles with 8, 824 unique authors (including authors from MSU and external institutions).

For each article, OpenAlex provides the article's title, abstract (if available), publisher, published year, digital object identifier (DOI), citation count, etc. Since we are interested in building a topic-oriented research network, we only considered an article's title and abstract for connecting researchers in the scholarly social network. However, we kept all of the harvested data in our database to enable additional features, for example, the past collaboration history and internal and external collaborators for Web visualization.

The OpenAlex harvest contains publications by researchers that are not currently employed by MSU. Therefore, we mapped the current faculty at MSU with OpenAlex harvested data, which provides 9, 659 research articles with at least one of the authors being at MSU. For some of the faculty, we did not obtain any publication data, and for some, the number of publications is less than 4, which is insufficient to generate a reasonable list of topics. Therefore, we excluded them from the topic network, which left a total of 326 researchers with a maximum of 179 and an average of 29 publications per researcher in the topic network. Finally, we used a MySQL database to store the OpenAlex harvested data.

[1] https://openalex.org/

4 SCHOLARLY RECOMMENDER SYSTEM

This demonstration paper builds upon our previously published work; for an in-depth understanding of the methodology, please see [14]. The proposed framework comprises three modules: 1) topic modeling, which aims to identify latent topics within the publication corpus; 2) scholarly social network construction, based on topic similarity between researchers; and 3) network analysis and community detection to identify cross-domain scholar communities with shared topics of interest.

4.1 Topic Modeling

Since we are interested in building a scholarly social network based on the topics or concepts of interest to the researchers, the first step is to find those hidden topics from the researchers' published articles. We treated an article's title and abstract as a single document and the collection of all research articles extracted from OpenAlex as the corpus. We used Latent Dirichlet Allocation (LDA) [1], a generative probabilistic model for discovering latent topics. To train the LDA topic model, we followed the standard text preprocessing steps using Python's NLTK library [9]. Also, we filtered out the extreme words that appeared in more than 70% of the documents and less than two times in the whole corpus, which left us with 35, 468 unique words in the vocabulary set from the 16, 827 documents. For the LDA model, we utilized the LDA *mallet* model [16], which uses Gibbs Sampling to estimate the model's parameter. LDA requires the number of topics to be prespecified, and we used the *UMASS* coherence score [10] to select the best topic number of 400 topics.

4.2 Network Construction

To construct a scholarly social network, we need to define the entities and their relationships as a graph structure. Researchers act as entities, and for the relationships, we utilized their topic-based similarity obtained from the trained LDA model. To identify the topic probability distribution of each researcher, we combined the content of their publications, treated it as a single document, and queried our trained LDA model to obtain the topic probability distribution of that document. The distribution will differ for each researcher based on their research fields and topics of interest. Next, to measure the similarity score between researchers to define the edge weight in the constructed network, we utilized the Jensen Shannon Divergence (JSD) score, which is a bounded and symmetric measure based on Kullback-Leibler (KL) divergence [17] that measures the similarity between two probability distributions. JSD is bound between 0 and 1, where values close to 0 indicate strong similarity. We used $1 - JSD$ for the edge weights between researchers to have larger weights correspond to strong similarity.

Constructing a network based on topic similarity results in a fully connected network, where each researcher is linked to every other researcher due to the $(1 - JSD)$ score always yielding a value greater than zero. However, network analysis on such a fully connected network can often present challenges in delivering

meaningful insights. To address this, we implemented an edge-weakening technique to refine the network structure, intending to enhance community analysis. We utilized various edge threshold values, ranging from 0.01 to 0.5, to weaken the network structure. Completely removing edges based on these threshold values might result in the formation of disconnected sub-networks at higher thresholds, rendering community analysis unsuitable. In response to this concern, rather than removing edges entirely, we chose to preserve edges in the network with a low edge weight of 0.0001.

4.3 Community Analysis

We utilized two discrete community detection algorithms, Louvain [2] and Spectral Clustering [13], capable of detecting communities in weighted graphs. We constructed networks based on different edge threshold values, and to evaluate the discovered communities, we utilized the modularity score [11], a measure to assess how well the members are grouped within the network.

The Louvain algorithm greedily optimizes the modularity score to discover communities and returns communities with the best modularity score. On the other hand, Spectral Clustering requires the number of communities to be prespecified, and as we are comparing the detected communities of these two algorithms, we set the number of communities in Spectral Clustering based on the result of the Louvain algorithm.

To evaluate the performance of the two community detection algorithms, we used the Jaccard similarity score between pairs of sets of nodes in the discovered communities. [5]. If members of both communities are identical, the similarity score would be one, and if they are completely different, it would be zero.

As both the algorithms are unsupervised, meaning the discovered community numbers may not align, we calculated all pairwise community Jaccard distances between these two algorithms and sorted the distance to obtain the community alignment between two different algorithms. Finally, for an overall similarity score of the communities discovered by the algorithms, we averaged the Jaccard similarity of the best-aligned community pair with respect to the total number of communities. If both algorithms produce the same set of communities, the overall score would be one.

5 EVALUATION & DISCUSSION

Figure 1 shows the modularity score from the two algorithms with different edge threshold values. As we can see, the modularity score improves with the increase of the threshold values until up to a certain threshold and starts to drop around 0.5, which suggests that weakening the network structure provides an advantage for the algorithms to detect more refined community structures.

Figure 2 shows the number of communities discovered by the Louvain algorithm (the right Y axis), which we kept the same for the Spectral Clustering algorithm. It also shows the average Jaccard similarity (the left Y axis) that compares the similarity of the discovered communities by the two algorithms. For the number of communities, we notice an increase in community size after the threshold of 0.2 until 0.4 and then start to decrease. For average Jaccard similarity, where a score of one represents communities discovered by the algorithms are identical, we observe a similarity score of 0.99 at a threshold value of 0.36. Therefore, we selected

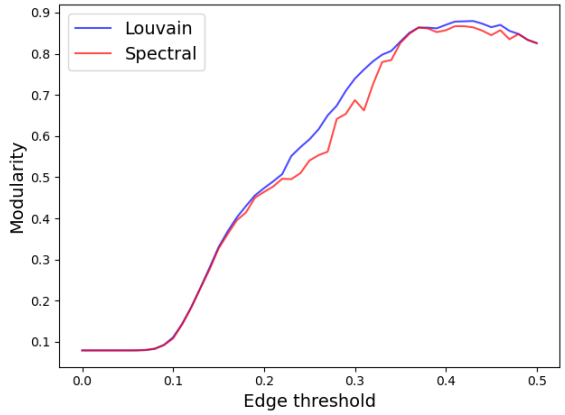

Figure 1: Modularity score based on different edge threshold

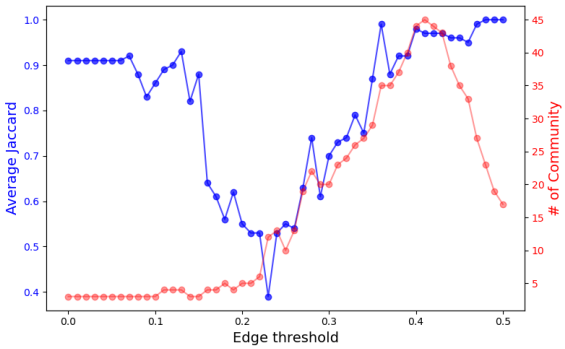

Figure 2: Average Jaccard and number of communities

the scholarly social network at a threshold of 0.36 to be our final network for cross-disciplinary community analysis. In this network, we obtained a total of 35 communities, with the largest community of size of 126, the smallest size of 2, and a median size of 3. The reason behind obtaining a large community of 126 members is that most edges get pruned in that threshold range and are clustered together into a big community. As we used a static threshold, many connections fell below the threshold, giving this large community. Using dynamic threshold or hierarchical community detection algorithms [12] may aid in breaking up big communities, which we left as possible future work.

In Figure 3, we show an example WordCloud for one of our discovered communities relevant to the specific community members. We also compared the discovered community members' edges with the past co-authorship record from OpenAlex to determine how many of them can be potential new collaborators along with the corresponding departments of the members to demonstrate the interdisciplinary aspects. For this community, there are a total of 31 internal edges where 6 of the connections co-authored before with 25 new potential connections. The 18 members are from departments Ecology, Animal Science, Environmental Science, and Earth Science. Finally, we utilized the PageRank [3] algorithm to recommend the top 10 potential collaborators for each researcher.

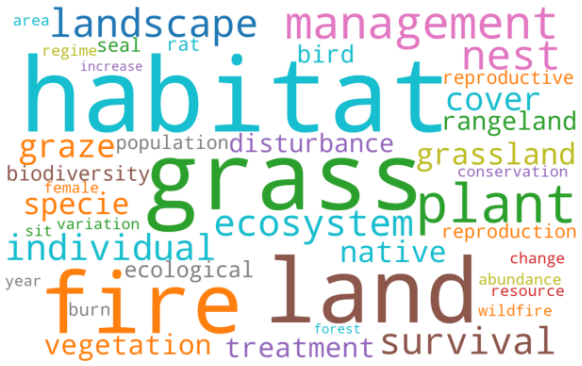

Figure 3: WordCloud of Community with 18 members

6 DEMONSTRATION & PATH TO IMPACT

Our ScholarNodes web application[2] includes a search screen, an individual researcher profile view, and a concepts browse view to show which community members are working within topics and domains. In the profile view, we created a force-directed graph drawing visualization identifying the interdisciplinary research community of a single researcher along with primary recommendations and other supplementary browse points including topic clusters and co-author networks. Recommendations are scored with the numerical values extracted from running the PageRank algorithm against a researcher's works and the nearest neighboring researchers and their works. The force-directed graph is interactive and allows for further exploration into related individual researchers. From this simple interface, a user can search and navigate to all the related nodes (researchers) in the community (Figure 4).

While our prototype is in an early phase, we have seen immediate impact within our campus community. The Center for Faculty Excellence and Office of Research Development have been regular users of the software to design faculty learning communities, match mentors for early-career faculty, and predict interdisciplinary grant teams. We anticipate continued use at the institutional level, but we also have grant work that will bring our implementation into the Montana University System (MUS) under the SMART FireS National Science Foundation grant. In this extension of our work, we are applying the clustering and searching prototype to recommend research collaborations external to the MUS SMART FireS grant team from entities such as NASA. We can carry out this work using additional datasets and connecting to our network analysis, clustering, and retrieval methods. We have also tested our partition and similarity algorithms on another long-form narrative, content-based dataset of student retention data in a project for the Office of Student Success. We were able to find and match communities in this dataset including: communities of support between students and students in need of support to continue matriculation. Given these early results, we have already demonstrated that our system is extensible when given new data and adaptable to other social network communities which speaks to potential reach and impact. And finally, we know there is a market for RIMS in higher education. We see our software, hybrid algorithmic practices, and open data routine having a longer-term impact as an option for university analytics and data offices looking to diagnose, forecast, and recommend scholarly relationships for their campus community.

[2]ScholarNodes is available at https://www.lib.montana.edu/msu-research-community.

7 SUMMARY & FUTURE WORK

Detecting communities in content-based social network datasets through network analysis remains a rich and open research thread. To this end, we presented ScholarNodes, a web interface designed for visualizing scholarly social networks and identifying interdisciplinary communities through the integration of content-based filtering and social network analysis. The analysis exhibited encouraging results in identifying cross-domain collaboration networks and identifying potential collaborators with diverse fields of study. The implications of our collaboration system extend beyond academia, resonating with the broader landscape of content-based recommendation systems.

To date, we have focused exclusively on discrete community detection algorithms, wherein a member is assigned to a single community. Moving forward, we aim to delve into overlapping or hierarchical community detection algorithms to enhance the performance of our community detection. Currently, our constructed network is static; however, in future research, we intend to explore the dynamic nature of the network to enhance the accuracy of recommendations. Additionally, we plan to expand our researcher dataset by including researchers from across the Montana University System's various institutions, thereby enabling cross-institutional analysis in our recommendation software.

ACKNOWLEDGMENTS

We would like to thank and recognize James Espeland, software engineer at the MSU Library, for his work to prototype the ScholarNodes data retrieval model and consult on OpenAlex data harvesting routines. This paper is based on work supported, in part, by NSF EPSCoR Cooperative Agreement OIA-2242802. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

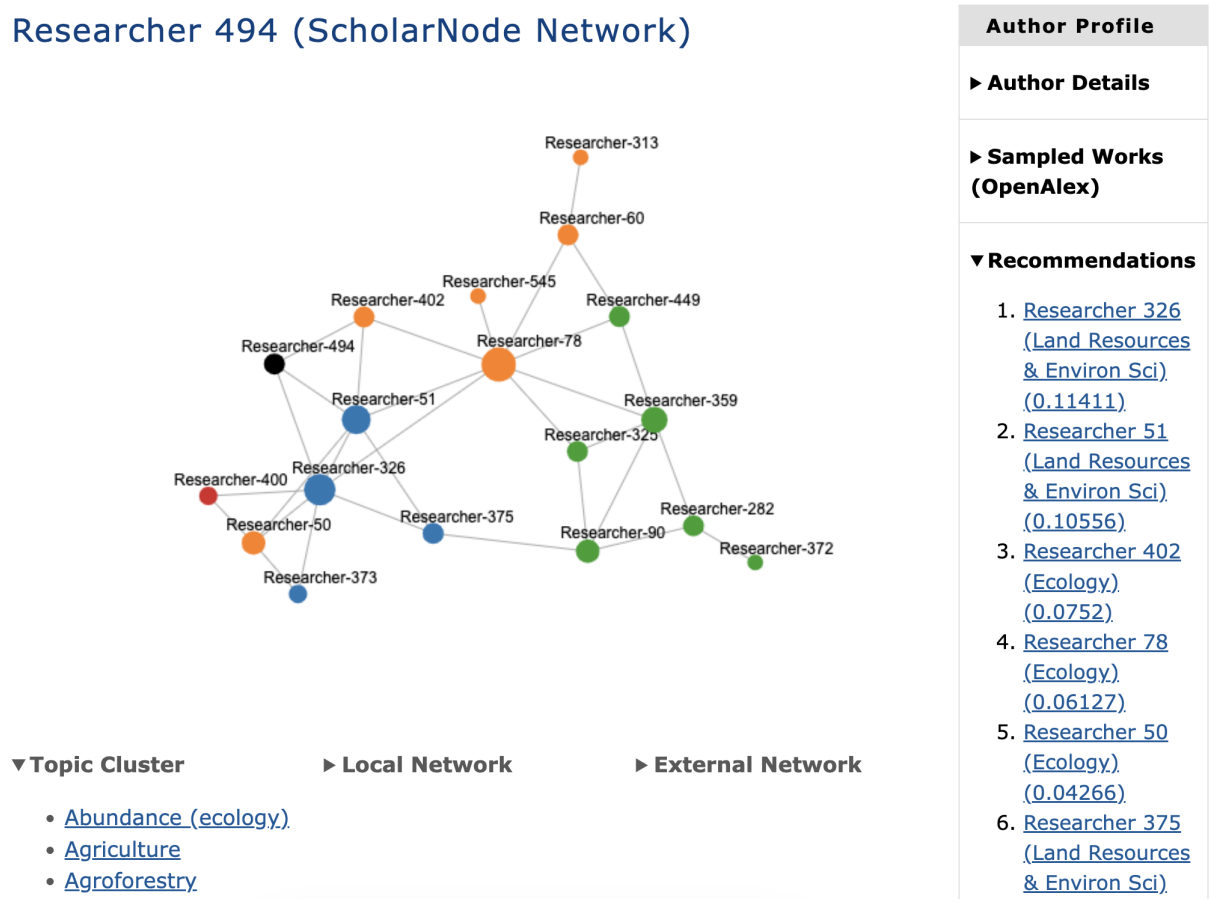

Figure 4: Profile view of scholar network & recommendations

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[2] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.

[3] Sergey Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30, 1-7 (1998), 107–117.

[4] Rebecca Bryant, Jan Fransen, Pablo de Castro, Brenna Helmstutler, and David Scherer. 2021. *Research Information Management in the United States: Part 1—Findings and Recommendations.* OCLC Research, Dublin, OH. https://doi.org/10.25333/8hgy-s428

[5] Paul Jaccard. 1912. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* 49 (1912), 547–579.

[6] Xiangjie Kong, Huizhen Jiang, Wei Wang, Teshome Megersa Bekele, Zhenzhen Xu, and Meng Wang. 2017. Exploring dynamic research interest and academic influence for scientific collaborator recommendation. *Scientometrics* 113 (2017), 369–385.

[7] Xiangjie Kong, Huizhen Jiang, Zhuo Yang, Zhenzhen Xu, Feng Xia, and Amr Tolba. 2016. Exploiting publication contents and collaboration networks for collaborator recommendation. *Public Library of Science* 11, 2 (2016), e0148492.

[8] Wei Liang, Xiaokang Zhou, Suzhen Huang, Chunhua Hu, and Qun Jin. 2017. Recommendation for cross-disciplinary collaboration based on potential research field discovery. In *2017 fifth international conference on advanced cloud and big data (CBD)*. 349–354.

[9] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. https://arxiv.org/abs/cs/0205028

[10] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 262–272.

[11] Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103, 23 (2006), 8577–8582.

[12] M. E. J. Newman. 2006. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 103, 23 (2006), 8577–8582.

[13] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2002. Spectral clustering for image segmentation. In *Proceedings of the 2002 conference on Advances in neural information processing systems*. 849–856.

[14] Md Asaduzzaman Noor, John Sheppard, and Jason Clark. 2023. Finding Potential Research Collaborations from Social Networks Derived from Topic Models. In *2023 10th International Conference on Behavioural and Social Computing (BESC)*. IEEE, 1–7.

[15] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833 [cs.DL] https://arxiv.org/abs/2205.01833

[16] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45–50.

[17] Sheldon M. Ross. 1997. *Introduction to Probability Models* (sixth ed.). Academic Press, San Diego, CA, USA.

[18] Xiaokang Zhou, Wei Liang, Kevin I-Kai Wang, Runhe Huang, and Qun Jin. 2021. Academic Influence Aware and Multidimensional Network Analysis for Research Collaboration Navigation Based on Scholarly Big Data. *IEEE Transactions on Emerging Topics in Computing* 9, 1 (2021), 246–257.

Received 08 February 2024; revised xx xxxxxx xxxx; accepted xx xxxxx xxxx