# Handling Publication Imbalance for Effective Community Detection in Scholarly Networks

Md Asaduzzaman Noor, John Sheppard, and Jason Clark

Montana State University, Bozeman MT-59717, USA
mdasaduzzamannoor@montana.edu, john.sheppard@montana.edu, and
jaclark@montana.edu

**Abstract.** Finding potential research collaborators is a challenging task, especially in today's fast-growing, interdisciplinary research landscape. While traditional methods rely on observable ties like co-authorships and citations, we focus solely on publication content to build a topic-based research network using BERTopic with a fine-tuned SciBERT model that connects and recommends researchers across disciplines based on shared topical interests. A key challenge we address is publication imbalance, where some researchers publish much more than others, often across several topics. Without careful handling, their less frequent interests are hidden under dominant topics, limiting the network's ability to capture their full research scope. To tackle this, we introduce a cloning strategy that clusters a researcher's publications and treats each cluster as a separate node. This allows researchers to belong to multiple communities, improving the detection of interdisciplinary links. Evaluation shows that the cloned network leads to more meaningful communities and uncovers broader collaboration opportunities.

**Keywords:** Community detection · Collaboration recommendation · Topic-based scholarly network · BERTopic · Social network analysis

## 1 Introduction

Research collaboration plays a crucial role in advancing scientific discovery, often leading to impactful and interdisciplinary outcomes. As the volume of scholarly publications grows, recommending meaningful collaborations has become increasingly challenging. Most existing approaches rely on observable relationships, such as co-authorship or citation networks, which tend to reinforce known connections and overlook opportunities based on shared topical interests.

We argue that topical similarity, derived from publication content, is a powerful relation for identifying potential collaborations. Researchers working on similar themes may never have co-authored a paper or even be aware of each other's work. By focusing on what researchers publish, we can uncover hidden connections and recommend more diverse collaborations beyond disciplinary lines.

While prior work has used publication data, most focus narrowly on ranking candidates with limited interpretability. In contrast, social network analysis

(SNA) offers a broader view, revealing how researchers are organized, highlighting influential nodes, and providing community-driven insights through network structures.

In this work, we construct a scholarly network based on topic similarity, using publication titles and abstracts to group researchers by shared research themes. Building on our prior work [11,12], we now address the challenge of publication imbalance. High-output researchers often work across multiple topics, but their less frequent interests are overshadowed by dominant ones, limiting the network's ability to detect meaningful connections.

To address this, we introduce a cloning strategy that clusters a researcher's publications into distinct topical groups, creating multiple "clones" that participate in different communities. This improves the detection of diverse and interdisciplinary collaborations.

Our key contributions are as follows. We introduce a cloning-based strategy to address publication imbalance and improve community detection in topic-based research networks. Using BERTopic on publication titles and abstracts, we build a researcher similarity matrix that captures diverse research interests. We empirically show that our approach uncovers more meaningful collaboration opportunities.

## 2 Related Work

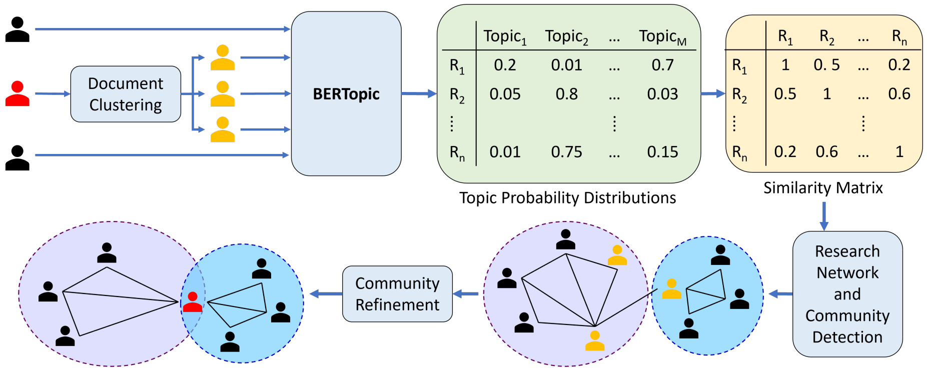
One key goal of researcher social network analysis is to recommend potential collaborators. Most existing work relies on direct relationships, such as co-authorship or citation links, to suggest collaborators or identify communities. While effective, these methods often overlook the topical diversity of individual researchers, limiting connections based on shared research interests.

Earlier studies framed collaboration recommendation as a link prediction problem on co-authorship networks [8,1]. Hybrid approaches combined direct links with content features to improve recommendations [15,6,16], but still depend on existing connections and tend to reinforce them rather than discover new interdisciplinary links.

Content-based methods rely solely on publication data to suggest collaborations. For example, Liang et al. [7] used LDA topic vectors for cross-disciplinary recommendations, and Kong et al. [5] modeled evolving interests with time-weighted topic distributions. However, these approaches focus on document similarity and top-$k$ recommendations, often lacking interpretability.

Integrating social network analysis with content data offers a more transparent and structural view, highlighting researcher clusters, topic distributions, and key participants through community detection.

Our work builds a topic-based researcher network while explicitly addressing publication count imbalance to enable fairer and more meaningful community detection. To our knowledge, this is the first study to tackle this issue in topic-based community detection for research collaboration.


Document Clustering
BERTopic
Topic Probability Distributions
Similarity Matrix
Research Network and Community Detection
Community Refinement

| | Topic$_1$ | Topic$_2$ | ... | Topic$_M$ |
|---|---|---|---|---|
| $R_1$ | 0.2 | 0.01 | ... | 0.7 |
| $R_2$ | 0.05 | 0.8 | ... | 0.03 |
| ⋮ | | | ⋮ | |
| $R_n$ | 0.01 | 0.75 | ... | 0.15 |

| | $R_1$ | $R_2$ | ... | $R_n$ |
|---|---|---|---|---|
| $R_1$ | 1 | 0. 5 | ... | 0.2 |
| $R_2$ | 0.5 | 1 | ... | 0.6 |
| ⋮ | | | ⋮ | |
| $R_n$ | 0.2 | 0.6 | ... | 1 |

Fig. 1: Overview of the proposed methodology

## 3 Dataset

To build the researcher dataset, we used Montana State University's (MSU) current faculty list and retrieved their publication history using OpenAlex [14], an open-source API for accessing scholarly metadata. We collected publication titles, abstracts, and author IDs from 2004 to 2025 for papers affiliated with MSU faculty. In total, we extracted metadata for $9,768$ publications. To ensure meaningful topic distributions for our network construction, we excluded researchers with fewer than five publications. This resulted in a final dataset of 296 faculty members, with a maximum of 190 publications for a single researcher and mean and median publication counts of 33 and 22, respectively.

The publication count distribution across researchers is heavily right-skewed: a small number of researchers publish significantly more than others. This imbalance highlights the need to address publication count disparities for effective community detection.

## 4 Methodology

Our proposed method (Figure 1) consists of several steps: training a topic model, cloning high-publication researchers, computing topic similarity, building the research network, detecting communities, and refining the structure.

**Topic Modeling.** We train a BERTopic model [4] on publication titles and abstracts from all researchers to extract topic distributions. To improve domain relevance, we fine-tune the sentence transformer using Masked Language Modeling (MLM) [3] on our corpus, enabling the model to better capture the context of scientific text specific to our dataset. The trained model produces topic-word distributions and document-topic probabilities that are used in the subsequent steps.

**Cloning High-Publication Researchers.** Researchers with a high publication count often work across diverse topics, but their dominant themes can overshadow less frequent ones in network construction. To address this, we clone researchers who have more than 1.5 times the median publication count, specifically those with over 33 publications in our dataset. We cluster their publications by first applying UMAP [10] for dimensionality reduction, followed by HDBSCAN [9] for document grouping. Each resulting cluster forms a "clone" that corresponds to a thematic area within the broader research landscape, allowing the researcher to participate in multiple communities aligned with their diverse interests.

**Computing Topic Similarity.** For each researcher or clone, we aggregate the topic probabilities of their publications to obtain a single topic distribution that represents their research focus. We then compute pairwise similarities using Jensen-Shannon Divergence (JSD), where lower divergence values indicate stronger topical alignment. This process results in a topic similarity matrix that captures the topical relationships between all researchers and their clones.

**Constructing the Research Network.** We construct a fully connected weighted graph using the similarity matrix as the adjacency matrix, where edge weights reflect topic similarity between researchers. To focus the network on meaningful connections, we prune edges with weights below a selected threshold, removing weaker links and highlighting more substantial topic alignments.

**Community Detection.** Communities are identified using the Nested Hierarchical Louvain (NH-Louvain) algorithm [13], which detects researcher groups at multiple levels of granularity. The hierarchical structure uncovered by this method naturally aligns with how research topics are organized, ranging from broad disciplines to more specific subfields.

**Refining Community Structure.** Since clones of the same researcher may appear multiple times within a single community, we refine the community structure by merging clones that belong to the same group. This ensures that each researcher is uniquely represented within a community while still preserving multi-community membership if their clones appear in different communities.

**Experimental Design and Hyperparameter Tuning.** We fine-tuned the sentence embeddings using MLM with 15% token masking over 40 epochs, starting from the `allenai/scibert_scivocab_uncased` model [2]. This fine-tuning step helped the model adapt to our research domain and improved the coherence of the extracted topics. For dimensionality reduction in BERTopic, we used UMAP with `n_neighbors = 15`, `n_components = 5`, and `min_dist = 0.0`, which preserved local document relationships in a low-dimensional space. Topic clustering was performed using HDBSCAN with `min_cluster_size = 8`

Table 1: Summary statistics of the researchers with clones

| | |
|---|---|
| Total Number of Researchers | 296 |
| High-Impact Researchers (More than 33 Papers) | 96 |
| Researchers with Clones | 68 |
| Max clones for a researcher | 10 |
| Median clones per researcher | 3 |

and `min_samples = 4`, which produced 445 distinct topics after training the fine-tuned model.

To create clones, we clustered each high-output researcher's publications using the same fine-tuned sentence embeddings. We applied HDBSCAN with `min_cluster_size = 10` and `min_samples = 5` to form clone groups that captured diverse topical areas within an individual's work. For the class-based TF-IDF representation, we applied standard NLP preprocessing steps, including stopword removal, digit and punctuation filtering, and lemmatization, to ensure cleaner topic representations. Hyperparameters throughout the pipeline were selected through random search, with a focus on finding interpretable and stable clusters rather than pursuing exhaustive optimization.

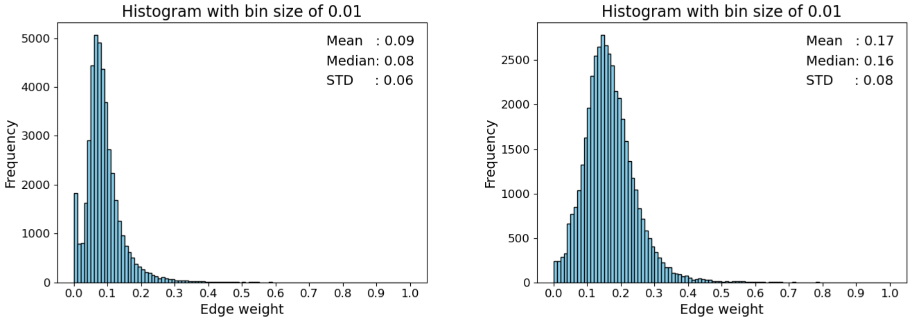
## 5 Results and Discussion

We begin by evaluating whether cloning improves community detection, especially for high-impact researchers. Although our method is unsupervised and lacks ground truth, we provide both quantitative and qualitative analyses to assess its impact. Table 1 summarizes the cloning outcomes. Of the 296 researchers in our dataset, 96 were classified as high-impact with more than 33 publications. Among them, 68 researchers formed multiple publication clusters via HDBSCAN, while the rest formed either a single cluster or were classified as outliers. The maximum number of clones for a researcher was 10, the median number of clones was 3, and the median publication count among clones was 23, which closely aligns with the original median of 22 before cloning.

We compared edge weight distributions before and after cloning, as shown in Figure 2. After applying the community refinement step to the cloned network, we observed reduced skew and increased mean and median edge weights, suggesting that topic similarity between researchers became more pronounced.

Figure 3 shows the mean edge weights of cloned researchers before and after cloning. For all cloned researchers, mean edge weights increased, indicating that cloning better captured their diverse research topics.

For community detection, we pruned edges until the network reached a density of 0.1. This threshold retains only strong connections, which helps uncover more meaningful and well-separated community structures. We used the NH-Louvain algorithm with a minimum community size of 30, balancing granularity and interpretability. The final network had 30 communities, with sizes ranging



(a) Before cloning (b) After cloning

Fig. 2: Distribution of edge weights before and after cloning

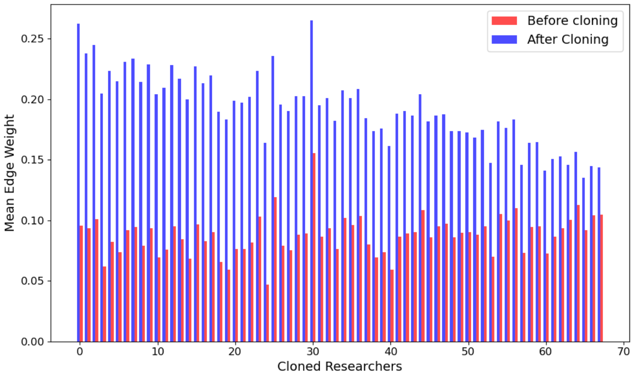


Fig. 3: Mean edge weights of cloned researchers before and after cloning

from 2 to 28 researchers (mean 11.2), and an average community density of 0.64, indicating strong internal connectivity.

We also observed overlapping community memberships. In total, 29 researchers belonged to more than one community after the refinement step, confirming that cloning helped reveal diverse topical affiliations. Most overlapping researchers appeared in two communities, with the maximum being four.

Figure 4 shows an example subnetwork where researchers 74, 228, 394, and 454 belong to multiple communities. Researcher 74 appears in all three communities, while others overlap between two.

Figure 5 provides a wordcloud example for overlapping researcher 454. The pre-cloning wordcloud (left) shows a mixture of topics, while the two clones (middle and right) show distinct topical focuses, allowing them to be placed in different communities.

Although our method was not explicitly designed for overlapping community detection, cloning naturally enabled multiple memberships. This opens the possi-

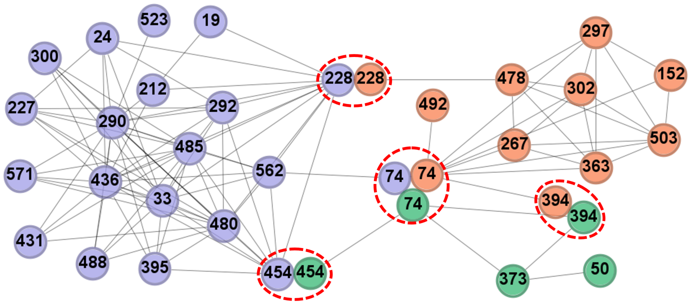
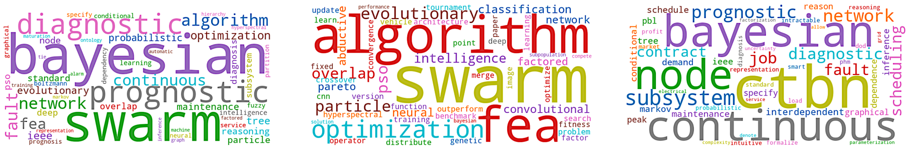


Fig. 4: Example of overlapping communities. Dashed circles indicate overlapping researchers.



(a) Researcher 454 (b) Clone 454-1 (c) Clone 454-2

Fig. 5: Wordclouds showing top 50 words for Researcher 454 and its two clones

bility of using our approach for detecting overlapping communities in text-based networks. Future work should also explore broader datasets, stronger baselines, sensitivity analyses on key thresholds, and scalability assessments to further validate and generalize the approach.

## 6 Conclusion

In this paper, we presented an approach exploring topic-based networks for identifying research communities and enabling diverse collaboration recommendations. Using BERTopic and a fine-tuned SciBERT model, we built a topic similarity network capturing connections across disciplines. To address publication imbalance, we introduced a cloning strategy that clusters publications, highlighting less common research areas otherwise overshadowed. This approach allows researchers to belong to multiple communities, better reflecting their full research scope and supporting interdisciplinary recommendations. Our evaluation shows that the cloned network improves both community coherence and the diversity of potential collaborations.

**Acknowledgments** 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.

## References

1. Backstrom, L., Leskovec, J.: Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 635–644 (2011)
2. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3615–3620 (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2019)
4. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)
5. Kong, X., Jiang, H., Wang, W., Bekele, T.M., Xu, Z., Wang, M.: Exploring dynamic research interest and academic influence for scientific collaborator recommendation. Scientometrics **113**, 369–385 (2017)
6. Kong, X., Jiang, H., Yang, Z., Xu, Z., Xia, F., Tolba, A.: Exploiting publication contents and collaboration networks for collaborator recommendation. Public Library of Science **11**(2), e0148492 (2016)
7. Liang, W., Zhou, X., Huang, S., Hu, C., Jin, Q.: Recommendation for cross-disciplinary collaboration based on potential research field discovery. In: 2017 fifth international conference on advanced cloud and big data (CBD). pp. 349–354 (2017)
8. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the twelfth international conference on Information and knowledge management. pp. 556–559 (2003)
9. McInnes, L., Healy, J., Astels, S.: hdbscan: Hierarchical density based clustering. Journal of Open Source Software **2**(11), 205 (2017)
10. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software **3**(29), 861 (2018)
11. Noor, M.A., Clark, J.A., Sheppard, J.W.: Scholarnodes: Applying content-based filtering to recommend interdisciplinary communities within scholarly social networks. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2791–2795 (2024)
12. Noor, M.A., Sheppard, J., Clark, J.: Finding potential research collaborations from social networks derived from topic models. In: 10th International Conference on Behavioural and Social Computing. pp. 1–7 (2023)
13. Noor, M.A., Sheppard, J.W., A. Clark, J.: Identifying hierarchical community structures in content-based scholarly social networks. In: 2024 International Conference on Machine Learning and Applications (ICMLA). pp. 440–447 (2024)
14. Priem, J., Piwowar, H., Orr, R.: Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833 (2022)
15. Yang, C., Sun, J., Ma, J., Zhang, S., Wang, G., Hua, Z.: Scientific collaborator recommendation in heterogeneous bibliographic networks. In: 2015 48th Hawaii International Conference on System Sciences. pp. 552–561 (2015)
16. Zhou, X., Liang, W., Wang, K.I.K., Huang, R., Jin, Q.: 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), 246–257 (2021)