Finding Potential Research Collaborations from Social Networks Derived from Topic Models

Md Asaduzzaman Noor
Gianforte School of Computing
Montana State University
mdasaduzzamannoor@montana.edu

John Sheppard
Gianforte School of Computing
Montana State University
john.sheppard@montana.edu

Jason Clark
MSU Library
Montana State Univeristy
jaclark@montana.edu

***Abstract*—Community detection is a valuable tool for analyzing social networks given its potential for identifying groups with common characteristics and common interests. In this work, we focused on detecting scholarly communities based on researchers' publication data to discover interdisciplinary collaboration recommendations (researchers working on different domains). Specifically, instead of using any physical or direct relationship between researchers, we utilized a topic model to obtain the topic-based similarity between researchers to construct the social network graph. Next, we employed an edge-weakening procedure to alter the initially constructed network to uncover a more refined community structure. Two community detection algorithms, Louvain and Spectral clustering, were utilized to find the community structures in the modified network. The results of our experiments revealed the ability to discover possible research communities for both algorithms that were comparable, which suggests that our method has the potential for identifying hidden interdisciplinary research collaboration recommendations using topical relationships as the basis for building and analyzing the social network graph.**

***Index Terms*—Social Network Analysis, Community Detection, Topic Modeling, Scholar Network**

I. Introduction

With the rapid growth of digital information such as social media, blogs, reviews, and scientific publications, the need for analyzing and mining these data has gained a lot of attention in the past decades. Social network analysis (SNA) is a valuable tool for analyzing these data that can group or cluster entities (e.g., people) based on their similarity in interest. In SNA, these groups are often called communities. When constructing a social network, one can take several approaches. One approach is taking direct or physical relationships such as friends, colleagues, coauthorship, or communication between people. Another approach is taking both direct and similarity relationships that do not exist physically, for example, topic-based similarity or co-citation. Most works in the existing literature focused on direct relationships to construct the network and used the indirect relationships either as node attributes or to strengthen the edge weight between the entities.

In this paper, we focus on the social network analysis of scholarly data to identify communities of researchers who might benefit from collaboration. For constructing the social network, instead of considering a direct relationship (e.g., coauthorship and author citation), we considered a topic-based similarity relationship. One of our intentions for this work was to investigate developing a recommender system that would be able to provide interdisciplinary (researchers from different fields/domains) collaboration recommendations. When constructing a social network based on a direct relationship, we hypothesize that it would not be able to provide the best interdisciplinary recommendations as researchers from different domains may not be connected directly. Therefore, we focused solely on content or topic-based similarity for the network construction, where the intuition is that researchers independent of domains working on similar topics of interest might benefit from being connected. For the topic-based scholar network, we used researchers' publication metadata consisting of publication titles and abstracts. Then with the metadata, we trained a topic model for topic discovery and connected researchers based on topic distribution similarity. Finally, we applied two community detection algorithms to cluster the researchers based on similar topics of interest.

II. Related Work

Traditional community detection algorithms such as that of Grivan and Newman [1], greedy modularity optimization [2], fast modularity optimization [3], and a spectral algorithm [4] work best when detecting communities based on the graph's topological structure. However, analyzing this structure, which in most cases only considers the direct relationship between actors (e.g., colleagues, co-author, co-citation), may impact the community detection results in the presence of additional content-based data (e.g., tweets, social media posts, publication data) [5]. Therefore, alternative work has been done that considers both topological and content-based features to improve the detection of communities.

Rosen-Zvi et al. [6] introduced an author-topic model that extends Latent Dirichlet Allocation (LDA) [7] to include authorship information. Following that, Liu et al. [8] introduced the Topic-Link LDA model that unifies the graph structure and topic models using a Bayesian hierarchical approach. To address the issue of growing complex networks, Wang et al. [9] proposed a local community detection method considering both link and content data where content features were extracted using TF-IDF scores [10].

Yang et al. [11] proposed a discriminative approach for combining link and content for detecting communities, where they used a conditional model to introduce hidden variables for detecting vertex popularity and a discriminative model for content analysis based on community membership. Liu et al. [12] combined graph structure and content following a content propagation perspective, where they modeled the interactions between vertices with influence propagation and random walks. As mentioned earlier, these methods focus on direct relationships for the network construction and use the content to strengthen the vertex popularity/relationship. However, we are more interested in constructing the scholar network solely based on content/topic similarity for interdisciplinary collaboration recommendations.

Some works focus solely on content for community detection. Velardi et al. [13] proposed a novel content-based model for social network analysis where they were able to detect the emergent semantics of the social network domain and introduced a measure of concept similarity based on lexical chains of ontological and co-occurrence relations. Similar to our work, Zhao et al. [14] introduced a topic-oriented community detection approach by grouping all the social objects into topic clusters and utilizing link analysis to detect topical communities. Similarly, Nguyen et al. [15] utilized the Author-Recepient-Topic (ART) model that discovers trending topics and automatically labels them for better product marketing. These works focus on topics as the vertex of the constructed network and do community detection based on these topic vertices. In our work, we introduced a novel way of connecting researchers with the topic-based probability distribution derived from a topic model such as LDA.

III. Problem Statement

The problem that we attempted to address in this study is, given a group of researchers' publication metadata (title, abstract, and authorship), how can we identify the set of communities based only on the topic-based relationships to be able to provide potential research collaboration recommendations? As we are more interested in discovering interdisciplinary research collaborations, we did not consider any direct or physical relationships such as authorship that could introduce a bias towards researchers working in the same domain.

We hypothesize that a social network constructed with topic-based relationships will be able to discover research communities based on similar topics of interest where members of the communities may have different fields or departments, suggesting interdisciplinary collaboration recommendations.

IV. Dataset

In this paper, we used publication metadata of the existing researchers at our institution for constructing the social network. One of the goals of this paper is to detect scholar communities not based on direct relationships such as co-author or citation but on indirect relationships such as topic similarity. A subsequent goal will be to use the detected communities to discover relations that can provide recommendations for interdisciplinary work. We used OpenAlex [16], which is an open-source platform to access the comprehensive interconnected catalog of scholarly papers, authors, institutions, venues, and more. OpenAlex collects data from many sources, including CrossRef, PubMed, institutional, and discipline-specific repositories.

The OpenAlex API allows filtering the publication data based on a specific institution id, and we used our institution id to obtain research articles where at least one of the authors had an affiliation with the institution. The time frame we used to extract the data was from 2004 to the Present (i.e., 2023). Each research article data from OpenAlex includes the article title, abstract, list of authors, publisher, publication date, citation count, document identifier (DOI), etc.; however, for building a topic-oriented network, we considered only the title and abstract of the article. For the given time frame, we collected data for 583 researchers, and we were able to extract the researcher's name, their college (e.g., College of Engineering, College of Letters and Science), and their department (e.g., Electrical Engineering, Physics, Ecology) from the school database. From OpenAlex, we downloaded data for $10,285$ research articles, where at least one of the researchers was an author from our institution. For some of the researchers, we did not obtain any publication data and for some, the total number of publications was less than 5 which may not be enough to obtain a reasonable number of topics. Therefore, we excluded them from the topic network. This left a total of 335 researchers in the topic network with a maximum of 171 publications and an average of 30 publications. Figure 1 shows the graph for the numbers of publications per researcher in descending order.

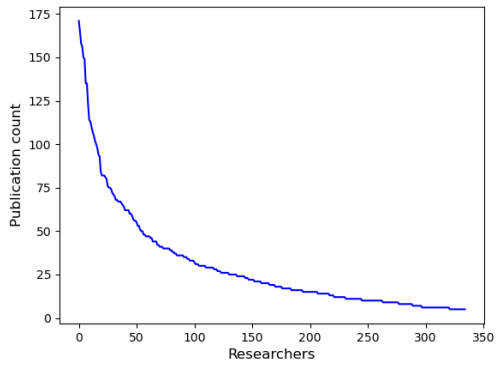


Fig. 1: Number of publications per researcher, sorted in descending order of count

V. Methodology

A. Topic Modeling

Since we are interested in building a scholarly network graph 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 used Latent Dirichlet Allocation (LDA) [7], a generative probabilistic model for discovering latent topics in a large corpus of documents, to uncover these topics. LDA treats documents as a mixture of latent topics, and topics as a mixture of words/terms seen in the documents, where documents are defined as a probability distribution over the latent topics sharing a common Dirichlet prior, and the topics are defined as a probability distribution over the words sharing a common Dirichlet prior as well.

In this work, 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. Then, we applied text preprocessing such as punctuation, digits, and stop-word removal to obtain the vocabulary set (words) for training. After training, we obtained the topic probability distribution of each researcher by querying the topic model with documents published by that specific researcher.

B. Constructing the Scholarly Social Network

A social network can be expressed as a graph structure $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where $\mathcal{G}$ represents the whole network, $\mathbf{V}$ represents the set of vertices, and $\mathbf{E}$ represents the set of edges. In our case, each vertex represents a researcher, and an edge between two researchers represents a relationship based on the topic similarity. Therefore, we need a measure or score to quantify the strength of the relationships to define the network edges. Recall that we extracted a topic probability distribution for each researcher using LDA.

We want the topic similarity score between the two researchers to be symmetric and bounded to do a fair network analysis later. Therefore, we used the Jensen-Shannon (JS) divergence, which is also based on the KL divergence, but it is symmetric, and the score is bounded between $[0, 1]$. The mathematical definition of JS divergence is

$$D_{JS}(\mathbf{P}||\mathbf{Q}) = \frac{1}{2}\left(D_{KL}(\mathbf{P}||\mathbf{M}) + D_{KL}(\mathbf{Q}||\mathbf{M})\right)$$

where P and Q represents two probability distributions, $D_{KL}(\mathbf{P}||\mathbf{Q}) = \sum_i P_i \log\left(\frac{P_i}{Q_i}\right)$, and $\mathbf{M} = \frac{1}{2}(\mathbf{P} + \mathbf{Q})$ is a mixed probability distribution. As the value of D_{JS} near zero indicates strong similarity, we used $1 - D_{JS}$ to define the edge relationship, which indicates that an edge weight near one represents a strong topic similarity between two researchers. Finally, by connecting all the researchers with their respective topic probability distribution, we ended up with an undirected weighted social network.

C. Community Detection

After constructing the social network, the next step is to do network analysis to discover meaningful patterns of information. Community detection is an effective tool for analyzing social networks, which from a graph perspective, is defined as a subset of vertices that are densely connected to each other and sparsely connected to the vertices in other communities in the same graph. Assume $\mathbf{C}_i$ is a community in a network $\mathcal{G}$ consisting of a subset of vertices $v_i \in \mathbf{V}$. The goal of community detection is to discover a community set $\mathcal{C} = \{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k\}$ such that $\mathbf{V} = \cup_{i=1}^{k} \mathbf{C}_i$. To evaluate the detected communities, we need a scoring function to tell us how good the found communities are.

Many scoring functions have been proposed in the literature to quantify the detected communities. Among them, the most popular is the Modularity metric [17], which is based on the intuition that sets of vertices that have many connections between their members should form a community. The modularity metric is defined as

$$Q = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(\mathbf{A}_{ij} - \frac{d_i d_j}{2m}\right) \times \delta(\mathbf{C}_i, \mathbf{C}_j)\right]$$

where $\mathbf{A}$ represents the adjacency matrix, $m = |\mathbf{E}|$ is the number of edges, d_i is the degree of vertex v_i, and $\delta(\cdot)$ is the Kronecker delta function (i.e., if vertex v_i and v_j are in the same community, it returns 1, otherwise 0). To summarize the equation, it measures if the fraction of edges inside a community is larger than the expected number of edges in the same community in a randomly initialized graph maintaining the original degree distribution. Alternative measures of evaluating communities are discussed in [18].

For this work, we used two community detection algorithms: Louvain and Spectral Clustering. The Louvain algorithm, introduced by Blondel et al. [3], is a greedy community detection algorithm that tries to partition the graph with a high modularity score. It follows an agglomerative approach, initially treating all the vertices as distinct communities. Then for each vertex, two calculations are performed: 1) compute the modularity gain (with respect to the whole graph) when putting the selected vertex to the community of its neighbor vertex, and 2) select the neighbor vertex that yields the highest modularity gain and merge the selected vertex community with the neighbor vertex community. The agglomeration continues until no further modularity gain is possible.

This process provides the first level of partition. In the second level, the algorithm makes a supervertex for each partition in the previous level and connects two supervertices by an edge if there is at least one edge connecting the partition in the previous level, and the weight of the edge is assigned as a sum the edge weights connecting the partition in the previous level. These two steps are then repeated to form new hierarchical levels and supergraphs until the communities become stable (i.e., no further gain in overall modularity score).

The Spectral Clustering community detection algorithm [4] relies on the spectral properties of a graph. If the communities are well-defined, then the rows across the eigenvectors of the graph Laplacian should be similar or close if they are in the same community. The symmetric graph Laplacian matrix, defined as $\mathbf{L}_{sym} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ where $\mathbf{D}$ is a diagonal degree matrix and $\mathbf{A}$ is the adjacency matrix of the graph, has some interesting properties such as the eigenvector decomposition of $\mathbf{L}_{sym}$ always provide real eigenvalues and the K largest eigenvalues corresponds to K cut in the graph. Then one can use any clustering algorithm such as K-means [19] by taking the eigenvectors corresponding to the K largest eigenvalues as input to cluster the vertices and assign communities.

VI. Experimental Design

For our approach, we needed to create the topic model based on the researchers' publication metadata (i.e., paper title and abstract). For the first step, we used Python's NTLK library [20] for digits, punctuations, stop-word removal, lemmatizing, and word tokenization for text-based data preprocessing. The next step was learning the LDA topic model for which we used Python's Gensim library [21], specifically the LDA mallet model that uses Gibbs sampling for learning the model's parameters. One of the hyperparameters for LDA is defining the number of topics beforehand. We used the topic coherence score [22] to determine the best number of K topics.

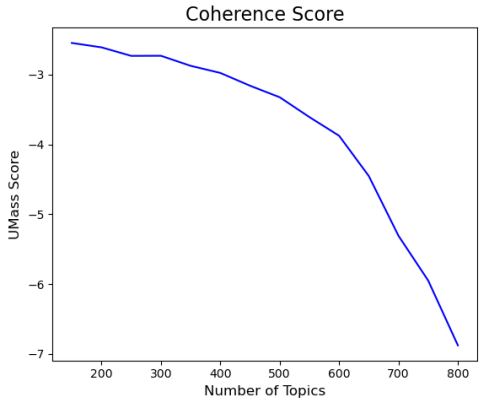


Fig. 2: UMass coherence score on different number of topics learned from collected research articles.

Specifically, the coherence score attempts to measure how interpretable the discovered topics are by taking the top N words of a topic and assessing how similar these words are across the corpus. We used the *UMass* [23] coherence score (the lower the better) defined as

$$C_{\text{UMass}}(w_i, w_j) = \log\left(\frac{D(w_i, w_j) + 1}{D(w_i)}\right)$$

where w_i and w_j correspond to two words, $D(w_i)$ counts how many times the word w_i appears alone in the corpus, and $D(w_i, w_j)$ counts how many times words (w_i, w_j) occur together in the documents. Finally, the global coherence score of a topic is the average pairwise coherence score on the top N words describing a topic. Figure 2 shows the *UMass* score on different numbers of topics in our experiments, and we selected the topic model with 600 topics (the elbow point) to build the social network graph.

Next, we calculated the topic probability distribution similarity, which was used to define the edge weight when connecting two researchers in the social network. Note that the JS divergence measure yields a value greater than zero even though the topic distribution may be dissimilar between two researchers. Thus, this process starts by creating a completely connected graph. Therefore, we tuned threshold values to weaken the edge weight, thus weakening the whole structure of the graph, to obtain a more refined community structure. Moreover, we used the modularity score as a measure to evaluate the detected communities by the different algorithms.

Spectral community detection needs the number of communities to be prespecified (a hyperparameter); however, the Louvain algorithm does not since it selects the number of communities with the highest modularity gain. As we compared the detected communities of these two algorithms, we set the number of communities in

Spectral clustering to be the same as obtained from the Louvain algorithm to make the results comparable. To evaluate the performance of the two community detection algorithms, we used the Jaccard similarity measure defined as

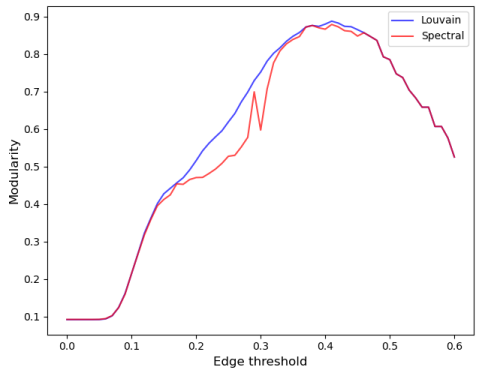
$$J(\mathbf{A}, \mathbf{B}) = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A} \cup \mathbf{B}|}$$

where in our case, $\mathbf{A}$ and $\mathbf{B}$ represent the set of vertices from two communities obtained by the two algorithms. If both $\mathbf{A}$ and $\mathbf{B}$ are identical, the similarity score would be one, and if they are completely different, the similarity score would be zero. As both the algorithms are unsupervised, meaning the discovered community numbers may not align, we calculated all pairwise community Jaccard distances and sorted the distance to obtain the community alignment between two different algorithms. Finally, to obtain an overall similarity score of the communities discovered by the two algorithms, we averaged the Jaccard similarity score 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.

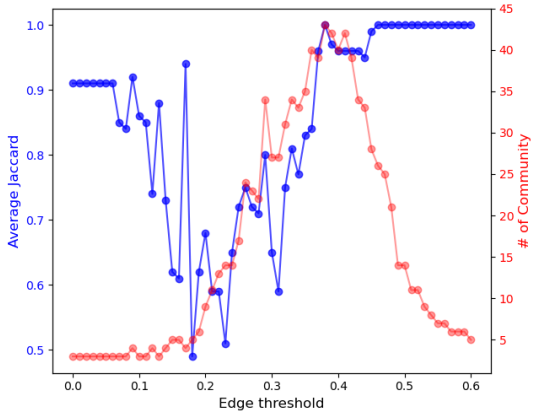
VII. Results & Discussion

To evaluate the effects of the edge threshold, we started with a threshold value of zero (i.e., the whole network was used without pruning) and then increased the threshold value in increments of 0.01 until 0.6. First, we applied the Louvain algorithm to detect the total number of communities, which was then used as the pre-specified number of clusters for the Spectral algorithm. Figure 3b (the right Y axis) shows the total number of communities discovered by the Louvain algorithm, and figure 3a shows the modularity score on the detected communities for different threshold values. For threshold values less than 0.15, both algorithms obtained a similar number of communities with a low modularity score. As the threshold increase, there is a gradual increase in the number of communities and modularity score for both algorithms, until threshold 0.42. At this point, communities and modularity gradually decrease. The modularity of the detected communities demonstrates that weakening the network structure is effective for obtaining a more refined network with a higher modularity score, at least to a point.

Next, we compared the similarity of the discovered communities by the two algorithms (Figure 3b), where an average Jaccard score of one implies that the detected communities of the algorithms are identical. For lower threshold values with fewer communities, we observe a high similarity score (0.9 on average) between the two algorithms, which then drops as the number of communities grows from a threshold value of 0.2 to 0.3 and gradually increases again afterward. At threshold



(a) Modularity score based on communities found.



(b) Jaccard similarity and number of communities between two algorithms.

Fig. 3: Characteristics of the detected communities based on different edge thresholds.

value 0.38, the average Jaccard similarity score becomes one, meaning the produced set of communities between the two algorithms is identical.

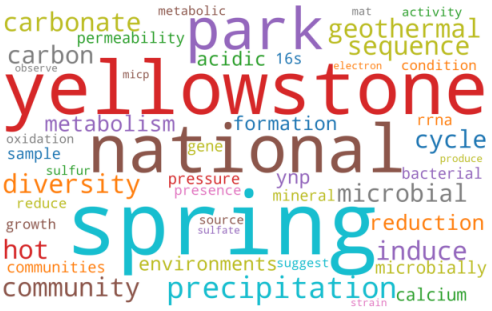
Table I shows the statistics of the discovered communities between the two algorithms where **TH**, **NC**, and **AvgJ** represent the threshold values, number of communities, and average Jaccard score respectively. The following columns list the maximum, minimum, median, and standard deviation of the sizes of the discovered communities with '**L**' and '**S**' specifying the Louvain and Spectral algorithm. The median community sizes gradually decrease as we increase the threshold values meaning most of the discovered communities have few members. After investigating the communities at threshold 0.38 (both algorithms produce the same set of communities), we noticed that there was only one large community with a size of 187, whereas the rest of the communities ranged from sizes 2 to 15.

After discovering the communities, qualitatively evaluated the communities by examining related

TABLE I: Louvain and Spectral Clustering Communities

Th	NC	AvgJ	Max(**L**) / Max(**S**)	Min(**L**) / Min(**S**)	Med(**L**) / Med(**S**)	Std(**L**) / Std(**S**)
0.00	3	0.91	140 / 136	78 / 77	117 / 122	31.34 / 30.83
0.06	3	0.91	141 / 136	79 / 77	115 / 122	31.13 / 30.83
0.12	4	0.74	117 / 121	48 / 53	85 / 80.5	29.15 / 28.09
0.18	5	0.49	102 / 125	30 / 12	77 / 62	29.69 / 40.27
0.24	14	0.65	70 / 120	2 / 3	14 / 18	22.08 / 30.25
0.30	27	0.65	100 / 102	2 / 2	5 / 3	19.50 / 22.88
0.36	40	0.84	157 / 151	2 / 2	2.5 / 3	24.54 / 23.61
0.42	39	0.96	233 / 235	2 / 2	2 / 2	36.90 / 37.22
0.48	21	1	289 / 289	2 / 2	2 / 2	62.57 / 62.57
0.54	8	1	320 / 320	2 / 2	2 / 2	112.38 / 112.38
0.60	5	1	327 / 327	2 / 2	2 / 2	145.34 / 145.34

(a) Communities with 13 members



(b) Communities with 9 members

Fig. 4: WordCloud of different communities

terms/vocabulary for each community. We anticipated the detected communities would align with the researchers' respective departments, given that researchers from similar departments should produce similar work (with some exceptions). Any exceptions are of particular interest and can be used to provide interdisciplinary collaboration recommendations. To identify the terms specific to a community, we combined all publications of the researchers of that community and queried the model to obtain the top five topics. Then we multiplied the topic probability by the term probability in that topic to identify the strengths of a term for that community.

Figure 4 shows example WordClouds for discovered communities at the threshold 0.38, chosen due to both algorithms producing identical sets of communities. We also aligned each community with the most represented department in that community. Subfigure 4a shows a community with size 13 where seven of the members are from Health & Human Development and the remaining members are from Psychology and Sociology & Anthropology. The terms in the WordCloud also show relevance to these departments. Subfigure 4b shows another example with a community size of nine members where there are six members (two members each) from Microbiology & Cell Biology, Environmental Science, and Chemical Engineering and the rest are from Civil Engineering. These examples demonstrate that the proposed model was able to identify communities with members of similar topic interests that are not necessarily from the same department.

VIII. Conclusion and Future Work

Based on our work constructing social networks from topic models from which communities were extracted, our results suggest that discovered communities align when using an appropriately tuned threshold to weaken the network structure. The extracted concepts for a community also provide valuable information such as members and concepts aligning with their respective departments and members from different departments working with similar concepts. The mixed membership communities are of interest to us as a way to provide interdisciplinary collaboration recommendations.

For this work, we only considered discrete community detection algorithms, meaning an entity can only belong to a single community. However, real-world datasets often contain entities that may belong to more than one community, referred to as overlapping communities. For future work, we will consider overlapping community detection algorithms to test if they improve the quality of the detected communities. In addition, we will investigate the impact of hierarchical and hidden communities. Furthermore, we will consider using different recommendation algorithms such as personalized PageRank [24] to obtain possible collaboration recommendations. Finally, we are applying our proposed model with different datasets, for example, the Beginning College Survey of Student Engagement Participation Agreement, which surveys entering college students' prior academic and co-curricular experiences, in the context of identifying undergraduate student cohorts as a means of improving student retention and reducing time-to-degree.

References

[1] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[2] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, p. 066111, 2004.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[4] L. Donetti and M. A. Muñoz, "Detecting network communities: a new systematic and efficient algorithm," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2004, no. 10, p. P10012, oct 2004.

[5] S. Ghiasifard, S. Khadivi, M. Asadpour, and A. Zafarian, "Improving the quality of overlapping community detection through link addition based on topic similarity," in *International Symposium on Artificial Intelligence and Signal Processing (AISP)*, 2015, pp. 182–187.

[6] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004, pp. 487–494.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[8] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link LDA: Joint models of topic and author community," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 665–672.

[9] C. Wang, W. Tang, Y. Wang, J. Fang, and S. Yao, "Local community detection algorithm based on links and content," in *IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2017, pp. 1805–1808.

[10] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[11] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 927–936.

[12] L. Liu, L. Xu, Z. Wangy, and E. Chen, "Community detection based on structure and content: A content propagation perspective," in *IEEE International Conference on Data Mining*, 2015, pp. 271–280.

[13] P. Velardi, R. Navigli, A. Cucchiarelli, and F. D'Antonio, "A new content-based model for social network analysis," in *IEEE International Conference on Semantic Computing*, 2008, pp. 18–25.

[14] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams, and J. Fan, "Topic oriented community detection through social objects and link analysis in social networks," *Knowledge-Based Systems*, vol. 26, pp. 164–173, 2012.

[15] M. Nguyen, T. Ho, and P. Do, "Social networks analysis based on topic modeling," in *International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, 2013, pp. 119–122.

[16] J. Priem, H. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," 2022. [Online]. Available: https://arxiv.org/abs/2205.01833

[17] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[18] F. D. Malliaros and M. Vazirgiannis, "Clustering and community detection in directed networks: A survey," *Physics Reports*, vol. 533, no. 4, pp. 95–142, 2013.

[19] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press, 1967, pp. 281–297.

[20] E. Loper and S. Bird, "Nltk: The natural language toolkit," 2002. [Online]. Available: https://arxiv.org/abs/cs/0205028

[21] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

[22] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[23] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, p. 952–961.

[24] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to

the web," Stanford InfoLab, Tech. Rep. SIDL-WP- 1999-0120, 1999.