CONTENT-BASED RECOMMENDATION VIA TOPIC MODELING AND SOCIAL NETWORK ANALYSIS

by

Md Asaduzzaman Noor

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2025

DEDICATION

To my late grandfather. Growing up in Bangladesh, he never had the opportunity to study much, but he always valued education deeply and celebrated every academic success we had. As kids, nothing made him happier than seeing us do well in school. I know he would have been genuinely proud to see me complete this doctorate. Even though he is not here, I carried his encouragement and hopes with me throughout this journey.

ACKNOWLEDGMENT

I would first like to thank my advisor, Dr. John Sheppard, for his support, guidance, and patience throughout my Ph.D. I am grateful for the chance to learn from him and truly honored to have been his student. His advice, feedback, and high expectations pushed me to grow academically and improved the quality of this work in many ways. I appreciate the opportunities and direction he gave me over the years.

I also want to thank Professor Jason Clark. He was the first one who introduced the idea that eventually grew into this dissertation, and he has been involved with this project from the very beginning. His help with research discussions, project support, and overall guidance made a major impact on this work. I appreciate everything he has done during my time in the program.

I am thankful to my committee members, Dr. Sean Yaw and Dr. Matthew Revelle, for their suggestions, feedback, and support throughout the dissertation process and during committee meetings.

I would like to thank my wife, Dr. Farshina Nazrul, for her constant support, patience, and encouragement. She listened to all my frustrations and worries, helped me stay motivated, and stood by me through the ups and downs of this journey. I could not have done this without her.

This work was supported in part by the NSF EPSCoR SMART FIRES program. I would also like to thank James Espeland from the MSU Library for his help in building the ScholarNode web prototype, and Professor Venice Bayrd for her support and feedback

ACKNOWLEDGMENT – CONTINUED

during our SMART FIRES project meetings.

I am grateful to my parents and siblings for their constant encouragement and belief in me. They have supported me at every stage of my life, and their trust in me has always been a source of strength. I would also like to thank my nephews and nieces in Bangladesh, whose calls and energy always lifted my mood during the tougher parts of this journey. I also want to thank my Numerical Intelligent Systems Laboratory (NISL) labmates for their discussions and suggestions over the years.

Finally, I acknowledge the use of AI-based writing tools to help with grammar, clarity, and editing during the writing process. All research ideas, methods, and contributions presented in this dissertation are entirely my own.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

Table Page

LIST OF TABLES – CONTINUED

Table Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF ALGORITHMS

Algorithm Page

ABSTRACT

The exponential growth of digital information and scholarly output has increased the need for intelligent systems that can identify relevant, interpretable, and equitable connections among entities. Traditional content-based recommender systems focus on item similarity but often overlook the relational structures that govern how knowledge and expertise are organized. This dissertation advances content-based recommendation by integrating topic modeling and social network analysis into a unified framework that represents semantic similarity as a network. The framework models relationships among entities through their topical proximity, enabling recommendations that are not only accurate but also transparent and structurally grounded.

The research is organized around four interrelated questions. The first investigates how topic modeling can be combined with network analysis to construct topic-based collaboration graphs that reveal latent research communities. By transforming document-level topic distributions into author-level profiles, this approach defines weighted network edges through topical similarity, producing networks that balance cohesion and diversity. The second question extends this representation to hierarchical community detection, introducing Nested Hierarchical Louvain (NH-Louvain) and Spectral Hierarchical Agglomerative Clustering (Spectral–HAC). These methods uncover multilevel community structures, allowing recommendations to operate at different granularities, from tightly focused collaborators within a subcommunity to broader interdisciplinary groups.

The third question addresses a broader data imbalance problem, demonstrated through the case of publication imbalance, where prolific authors dominate the content space and bias recommendation outcomes. A cloning-based strategy was developed to represent such authors by multiple topical instances, each reflecting a distinct research direction. Clone-LDA and Clone-BERT variants reduce dominance effects, improve thematic diversity, and enhance background representation in the generated networks. The fourth question evaluates the framework's accuracy, stability, and its content-based explainability, assessed through hold-out and perturbation experiments. Results show that topic-based similarity remains stable under missing information and that hierarchical and cloned models yield balanced, semantically coherent communities. To operationalize these findings, the ScholarNode prototype system was developed, providing an interactive, explainable interface that links recommendations to their underlying topical and community evidence.

Together, these contributions establish a principled foundation for topic-driven, network-aware recommender systems. The integrated framework advances understanding of how semantic and relational information interact, while the ScholarNode implementation shows its practical feasibility. Beyond the scholarly domain, the same design principles, representing content similarity as a network, detecting hierarchical communities, addressing imbalance, and supporting content-based explanations, can generalize to other content-rich environments. This research thus lays the groundwork for recommender systems that emphasize clarity, diversity, and balanced representation.

CHAPTER ONE

INTRODUCTION

The rapid growth of digital platforms has transformed how individuals access, generate, and share information. From e-commerce and entertainment to professional and educational services, users now interact with immense volumes of digital content daily. Platforms such as Amazon, YouTube, and LinkedIn continuously generate new data streams, which has intensified the challenge of identifying information that aligns with individual needs and preferences. This phenomenon, often referred to as the "information overload" problem [53], underscores the importance of systems capable of filtering, prioritizing, and recommending content efficiently.

Information Retrieval (IR) and recommender systems have become indispensable for navigating this abundance of data. IR systems operate on an explicit query-response paradigm, where users provide search terms and receive ranked results that best match their intent [6]. In contrast, recommender systems function more implicitly by observing user interactions, learning from patterns of preference, and predicting new items of potential interest [100]. Today, recommender systems drive personalization across diverse applications, including product suggestions in online marketplaces, playlist generation in music platforms, and friend or connection suggestions in social networks.

Recommender systems are generally categorized into collaborative filtering, content-based filtering, and hybrid approaches [2, 10]. Collaborative filtering (CF) techniques predict a user's interests by analyzing the behaviors of similar users, operating under the assumption that individuals who agreed in the past will likely agree in the future [99]. Although CF has demonstrated strong performance in domains with dense interaction data, it suffers from

several limitations, notably the "cold-start" and "popularity bias" problems [47]. CF systems rely heavily on historical user-item interactions, making them less effective when new users or items lack sufficient interaction data, and tend to reinforce already popular items while neglecting novel or niche options.

Content-based filtering (CBF) [91] offers an alternative paradigm by recommending items that share similar features with those a user previously liked. Here, the system models a user's profile through the attributes of consumed items and recommends new items whose content features exhibit high similarity. These attributes may include keywords, textual descriptions, or latent representations extracted from unstructured data. For instance, in text-based domains, CBF methods often employ Natural Language Processing (NLP) to encode documents into feature-based representations, often expressed as vectors, that support similarity computation [8]. In essence, a recommender system aims to predict the utility of an item i for a user u, denoted by $R(u, i)$, and recommend items with the highest estimated utility $\hat{R}(u, i)$. In practice, this is operationalized by selecting the top-k items with the largest predicted scores for each user. This formalization provides a conceptual foundation for comparing how different recommender models approximate the true utility function.

Despite their advantages, CBF approaches face their own challenges. They require sufficient and high-quality content data to derive meaningful item representations, a limitation known as the "limited content analysis" problem [110]. Computing pairwise similarities among large numbers of items or users can also become computationally intensive, posing scalability concerns. Furthermore, CBF systems often treat users as independent entities and ignore potential relationships among them, which restricts the system's ability to capture social or collaborative dimensions of relevance. As a result, recommendations can appear accurate at the content level but less meaningful in terms of real-world context or user connectivity.

Social Network Analysis (SNA) offers a valuable perspective for addressing these limitations by representing users and their interactions as vertices and edges within a graph structure. SNA facilitates the discovery of communities, influence patterns, and relational dynamics that underlie information flow [71]. Integrating SNA with recommender systems adds a relational layer that can enhance personalization and interpretability [73]. By incorporating structural information such as communities, proximity, and influence, recommender systems can move beyond isolated content similarity and consider how social connections affect information relevance and trust. For example, users connected within the same network cluster may share overlapping interests that pure content-based similarity fails to capture.

While several hybrid approaches combine CF and CBF methods to exploit complementary strengths [13], few systems explicitly incorporate network structures into content-based recommendation. This gap becomes particularly evident in domains where content itself reflects complex relationships, such as research publications, patents, or educational materials, where content similarity and relational structure jointly shape meaningful connections. The intersection of SNA and content-based recommendation thus provides an opportunity to design systems that are not only accurate but also explainable and socially coherent.

The growing availability of structured and unstructured data, combined with advances in machine learning and natural language understanding, has made such integration increasingly feasible. Topic modeling and embedding-based representations have proven effective for uncovering latent thematic patterns in large text corpora [98]. When combined with SNA, these representations can form the basis of content-driven networks where vertices represent entities and edges capture topical similarity. This dissertation builds upon that intersection by developing a framework that leverages topic modeling and network analysis to identify, evaluate, and explain potential connections among users or items. Although this

framework is evaluated within scholarly networks, its principles generalize to any domain where content serves as the primary basis for recommendation.

1.1 Motivation

The increasing heterogeneity of digital ecosystems requires recommendation mechanisms that can adapt to diverse data structures, account for uneven content distribution, and offer interpretable results. Traditional CF and CBF methods, while effective within homogeneous datasets, encounter difficulties when applied to complex, content-rich domains. For example, content sources may vary in length, quality, and thematic scope, and users may engage with multiple domains simultaneously. Such conditions lead to representation imbalance, reduced comparability, and diminished model explainability.

Network-based perspectives help mitigate these challenges by capturing how content similarities aggregate into higher-order structures such as communities or clusters. These structures provide contextual cues that can guide recommendations beyond mere pairwise similarity. However, most existing hybrid recommender systems use networks as auxiliary representations rather than as the primary analytical backbone. The motivation behind this dissertation is to treat the network itself as a first-class object of analysis, where the structure emerges directly from content similarity and topic relationships. By doing so, the recommendation process can exploit both textual semantics and network topology in a unified manner.

Another key motivation is the need for explanations in how recommendations are formed. Modern recommender systems increasingly employ deep neural or embedding-based representations that, while powerful, often operate as black boxes. In this dissertation, interpretability refers specifically to the ability to understand the topical and structural factors that drive similarity, rather than to broader notions of model interpretability in machine learning. Likewise, explainability is used in a content-based sense, meaning

that users can trace recommendations back to observable features such as shared topics, overlapping word distributions, or community structure. Without explainability, users may find it difficult to understand or trust the system's recommendations [135].

Topic modeling, when integrated with network structures, provides a natural pathway to explanation: both topics and communities can be visualized, interpreted, and traced back to observable content features. This dual-level representation allows researchers, practitioners, or end-users to comprehend why certain recommendations are made, bridging the gap between algorithmic output and human interpretability.

Scalability also remains a pressing concern. As data grows exponentially, the ability to maintain efficient and meaningful recommendations becomes increasingly challenging. Network-based clustering and hierarchical community detection can help reduce computational complexity by summarizing large content graphs into interpretable structures, enabling both localized and global analysis. These capabilities make the framework well suited for institutions or organizations that handle large and evolving collections of textual artifacts.

Finally, this study aims to address representational imbalance, a common but often overlooked issue in content-based systems. In scholarly or creative domains, certain entities, such as prolific authors or highly active contributors, dominate content generation, leading to skewed topic distributions. This imbalance can obscure secondary or interdisciplinary themes that are critical for discovery and innovation. By developing mechanisms to balance such representation, the framework seeks to ensure that recommendations remain equitable and diverse.

Although the proposed approach is broadly applicable to any content-based recommender system, this dissertation evaluates its performance within scholarly networks derived from three institutions and a combined multi-institution dataset to assess scalability and robustness. Scholarly data offers a rigorous and interpretable test environment, as it encapsulates well-defined textual content (publications), identifiable users (researchers), and

observable relationships (coauthorships). This setting provides both the complexity and clarity required to evaluate the proposed framework, while maintaining generalizability to other content-based recommendation contexts.

1.2 Research Question

This dissertation investigates how topic-based representations of scholarly output can be integrated with social network analysis to improve the identification of potential research collaborations. The study is guided by four research questions that address the construction, refinement, and evaluation of topic-based collaboration networks.

- **RQ1:** How can topic modeling be integrated with social network analysis to promote cross-domain collaboration recommendations?

  Modern scholarly collaboration networks often rely on coauthorship or citation-based relationships that overlook latent topical connections between researchers. This question investigates how topic modeling techniques can be used to construct similarity networks that reveal cross-domain research alignment and enable more inclusive collaboration recommendations.

- **RQ2:** How does modeling hierarchical communities affect recommendations and community quality?

  Traditional community detection methods identify flat, non-overlapping structures that fail to capture the nested nature of real scholarly groups. This question examines whether hierarchical community modeling can uncover finer-grained and multi-level research communities, thereby improving both the interpretability and precision of collaboration recommendations.

- **RQ3:** How does accounting for data imbalance affect community detection, and how can we mitigate its effect?

Institutional research data are typically skewed, with a small group of prolific authors contributing disproportionately to the publication corpus. This imbalance can distort topic distributions and community assignments. This question explores how such imbalance influences network structure and investigates mitigation strategies, such as cloning prolific researchers across their distinct topical domains, to preserve balanced and representative community structures.

- **RQ4:** How effective is the proposed collaboration recommendation framework, and how can we provide interpretable explanations for recommended connections in scholarly networks?

  Beyond structural modeling, a practical framework must produce verifiable and interpretable recommendations. This question evaluates the performance and stability of the proposed models using historical coauthorship as a proxy for ground truth and investigates how explanation mechanisms through topical overlap and community context can help users understand and trust recommended collaborations.

1.3 Contribution

The following contributions summarize the primary outcomes of this dissertation and how they address the research questions.

This dissertation makes five contributions. First, it introduces a topic-based framework that integrates content modeling with social network analysis to construct collaboration networks grounded in topical similarity. This framework provides a foundation for identifying meaningful research themes and potential cross-domain connections beyond those visible in coauthorship data.

Second, it develops hierarchical community modeling techniques that capture multi-level structure within scholarly networks. These methods provide finer-grained insights into

topical organization and support recommendations at multiple levels of abstraction.

Third, it examines the role of data imbalance in content-based networks, using publication imbalance as a representative case, and proposes a cloning-based strategy that improves representational balance and more accurately reflects the diverse topical profiles of prolific researchers.

Fourth, it evaluates the accuracy and stability of the framework through experiments based on historical coauthorship data. The findings show that topic-based similarity remains robust under missing information and that the framework offers transparent, content-grounded reasoning for recommended collaborations.

Finally, the dissertation presents ScholarNode, an interactive prototype that implements the proposed framework. ScholarNode integrates modeling, network construction, community analysis, and visualization into a unified system, demonstrating the practical feasibility of the methods developed in this research.

1.4 Dissertation Organization

The remainder of this dissertation is organized into eight chapters following this introduction. Chapter 2 presents the theoretical and methodological background necessary to situate the proposed framework. It introduces fundamental concepts of Social Network Analysis, including graph construction, edge weighting, and community detection algorithms, along with unsupervised evaluation measures such as modularity and conductance. The chapter also summarizes essential Natural Language Processing techniques used for textual preprocessing, representation, and topic extraction, including approaches for computing coherence scores to assess topic quality.

Chapter 3 describes the data collection and preparation procedures used throughout this study. It details the retrieval of publication metadata for three institutions, the subsequent filtering and cleaning processes, and the compilation of descriptive statistics

for each institutional corpus. The dataset chapter establishes the empirical foundation for all subsequent analyses and experiments.

Chapters 4 through 7 correspond to the four central research questions that guide this dissertation. Chapter 4 investigates how topic modeling can be integrated with network analysis to construct topic-based collaboration graphs and promote cross-domain recommendations. Chapter 5 explores hierarchical extensions of community detection to capture nested and multi-level scholarly structures that better reflect real-world interdisciplinary groupings. Chapter 6 examines the impact of publication imbalance on community detection and proposes cloning-based strategies to mitigate its effects within topic models. Chapter 7 evaluates the performance, stability, and interpretability of the proposed recommendation framework, analyzing its reliability under varying data conditions and cross-institution scenarios.

Chapter 8 presents the "ScholarNode" web application, a functional prototype that operationalizes the complete framework developed in this dissertation. The system integrates topic modeling, network construction, and community visualization within an interactive web environment. It provides researcher-level and community-level insights, offering interpretable recommendations through visual and textual explanations that connect directly to the underlying model components.

Finally, Chapter 9 concludes the dissertation by summarizing the major findings and methodological contributions, highlighting limitations, and identifying directions for future work. It also discusses the broader applicability of the proposed framework beyond scholarly collaboration, emphasizing its potential for other content-based recommendation domains such as organizational knowledge discovery, research–industry partnership identification, and educational content matching.

CHAPTER TWO

BACKGROUND

This chapter provides an overview of the foundational concepts underlying this dissertation. It begins by introducing the principles of Social Network Analysis (SNA), including graph representations, structural properties, and community structures. Next, it reviews key community detection methods and evaluation metrics commonly used to assess community quality. The latter part of the chapter focuses on text analytics and topic modeling techniques, which serve as the basis for constructing content-based scholarly networks in subsequent chapters.

2.1 Social Network Analysis

We can think of a social network as a set of entities and the social relationships (e.g., family, friendship, colleagues) between them. A convenient way to represent a social network is as a graph, where vertices represent the entities and edges represent the interactions or relationships. Formally, we define a social network as a graph $\mathcal{G} = (V, E)$, where $\mathcal{G}$ is the entire network, V is the set of vertices, and E is the set of edges.

Importantly, the entities in a social network are not limited to people. From a marketing perspective, the entities could be products, and the relationships could represent how often those products are purchased together. In transportation, the entities could be hubs such as airports or bus stations, with relationships representing routes connecting them. Much like online platforms such as Facebook and Twitter, social networks and their analysis have become valuable tools in research, healthcare, business, environmental science, and many other fields [126]. SNA allows researchers to uncover hidden patterns, predict outcomes, and extract meaningful insights from these complex networks.

In a social network, relationships between entities can be either undirected or directed. A co-authorship network, where vertices represent researchers and edges represent joint publications, is undirected since the relationship holds in both directions. A citation network, on the other hand, is directed: if researcher A cites researcher B, the edge points from A to B but not necessarily the other way around.

Relationships can also be unweighted or weighted. An unweighted relationship simply indicates that a relationship exists between two entities. For example, in a friendship network, two vertices may be connected if the individuals are friends, without specifying the strength of the friendship. In contrast, a co-authorship network often uses weighted edges, where the weight corresponds to the number of papers the two researchers have co-authored.

Social networks can further be classified as homogeneous or heterogeneous [134]. In homogeneous networks, all connections between entities are of the same type, whereas heterogeneous networks contain multiple types of entities or relationships. For example, a friendship network is homogeneous, while an academic scholar network can be heterogeneous when it incorporates multiple relationship types such as coauthorship, citation, or institutional affiliation. An illustration of such a heterogeneous scholarly network is shown in Figure 2.1.

Relationships in social networks can also be direct or indirect. Direct relationships can be observed explicitly. Friendship, co-author, or citation links represent clear and immediate interactions between entities. Indirect relationships, on the other hand, are not observed directly but inferred through intermediate information. For instance, two researchers who have never collaborated might still be connected if they work on similar topics. These inferred links are critical in many SNA tasks, such as recommendation systems, content discovery, and community analysis, because they reveal shared interests and hidden structures within the network [128].

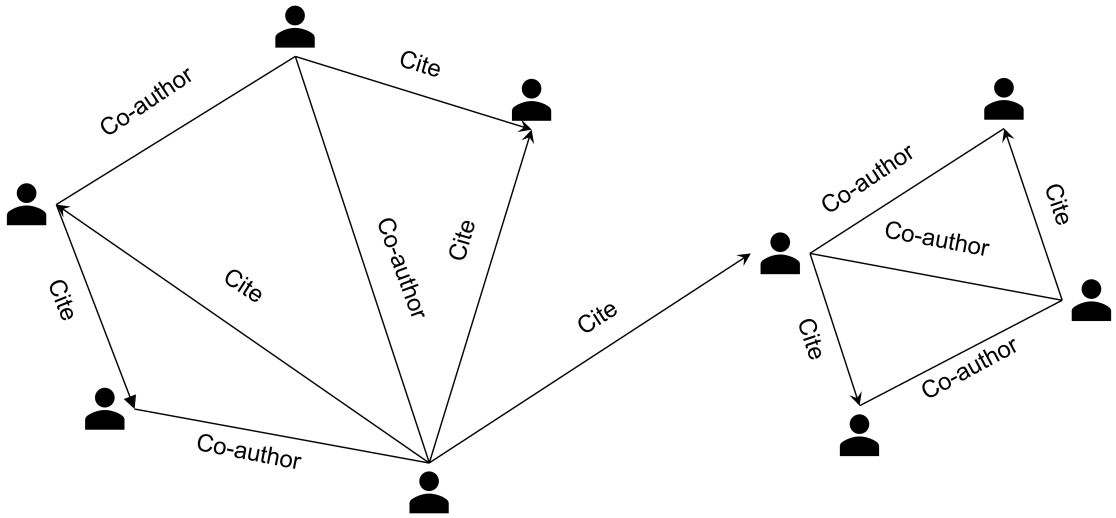


Figure 2.1: Heterogeneous academic social network

Finally, many real-world social networks include attributes associated with entities and their relationships. These attributes provide additional context and enable richer analysis. For example, in a friendship network, attributes like age or hobbies can reveal whether people with similar characteristics are more likely to connect. In an academic co-authorship network, researchers may have attributes such as academic rank (e.g., professor, graduate student), research area (e.g., AI, data mining), or institutional affiliation. Incorporating such attributes allows deeper insights, for example by identifying influential researchers by rank, clustering researchers with similar interests, or examining intra- and inter-institutional collaboration patterns. As a result, attributed SNA is widely applied across sociology, marketing, healthcare, and online social networks [16].

2.1.1 Structural Properties of Social Network

Understanding the structure of a social network often relies on fundamental concepts from graph theory. These concepts provide a mathematical framework to represent networks, quantify their characteristics, and analyze how entities interact. In this subsection, we

review key structural properties and measures that are commonly used in Social Network Analysis (SNA), including network representation, vertex degree, network density, and several centrality measures.

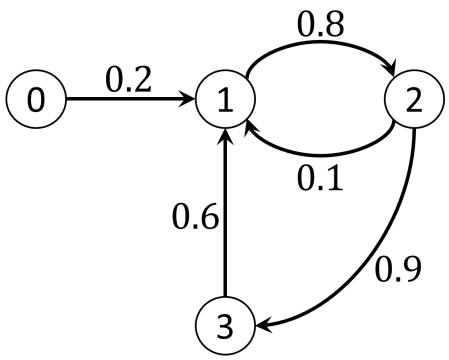
- **Network Representation.** A graph $\mathcal{G} = (V, E)$ is used to represent a network structure, where $V = \{1, 2, \ldots, n\}$ represents the set of vertices (actors or entities in the network) and $E = \{(i, j) \mid i, j \in V\}$ represents the set of edges. A pair (i, j) belongs to E if there is an interaction between actor i and actor j, and the size of the edge set is $|E| = m$. The mathematical representation for an attributed social network is $\mathcal{G} = (V, E, X)$, where X indicates the set of vertex attribute vectors containing additional information about each vertex (e.g., demographic details, professional roles, or affiliations).

  A graph is often represented by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ where

$$\mathbf{A}_{i,j} = \begin{cases} 1 & \text{if } (i, j) \in E, \\ 0 & \text{otherwise,} \end{cases}$$

  in the case of unweighted edges. For weighted graphs, $\mathbf{A}_{i,j}$ takes the value of the edge weight. Representing the network as a matrix is advantageous because it enables analysis using tools from matrix theory and linear algebra, which are closely connected to spectral graph analysis [17]. Figure 2.2 shows an example of a directed network with its adjacency matrix representation.

- **Degree.** The degree of a vertex is the number of connections it has. For a directed graph, it is further divided into in-degree and out-degree. In-degree represents the number of incoming connections a vertex has, and out-degree represents the number of outgoing connections of that vertex. In Figure 2.2, vertex 1 has an in-degree of 3 and



$$A = \begin{bmatrix} 0.0 & 0.2 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.0 \\ 0.0 & 0.1 & 0.0 & 0.9 \\ 0.0 & 0.6 & 0.0 & 0.0 \end{bmatrix}$$

Figure 2.2: Weighted directed network and adjacency matrix

an out-degree of 1. The vertices connected to a given vertex are called the neighbors of that vertex.

For an unweighted graph, the degree of a vertex v_i, denoted d_i, can be computed using the adjacency matrix $\mathbf{A}$ as:

$$d_i = \sum_{j=1}^{n} \mathbf{A}_{ij}$$

where $\mathbf{A}_{ij} = 1$ if there is an edge between vertex v_i and vertex v_j, and $\mathbf{A}_{ij} = 0$ otherwise.

For a weighted graph, the weighted degree of a vertex v_i, denoted w_i, is calculated as:

$$w_i = \sum_{j=1}^{n} \mathbf{W}_{ij}$$

where $\mathbf{W}_{ij}$ represents the weight of the edge between v_i and v_j (or 0 if there is no edge).

- **Density.** The density of a network quantifies the ratio of the number of actual connections to the total possible connections, providing insight into how interconnected

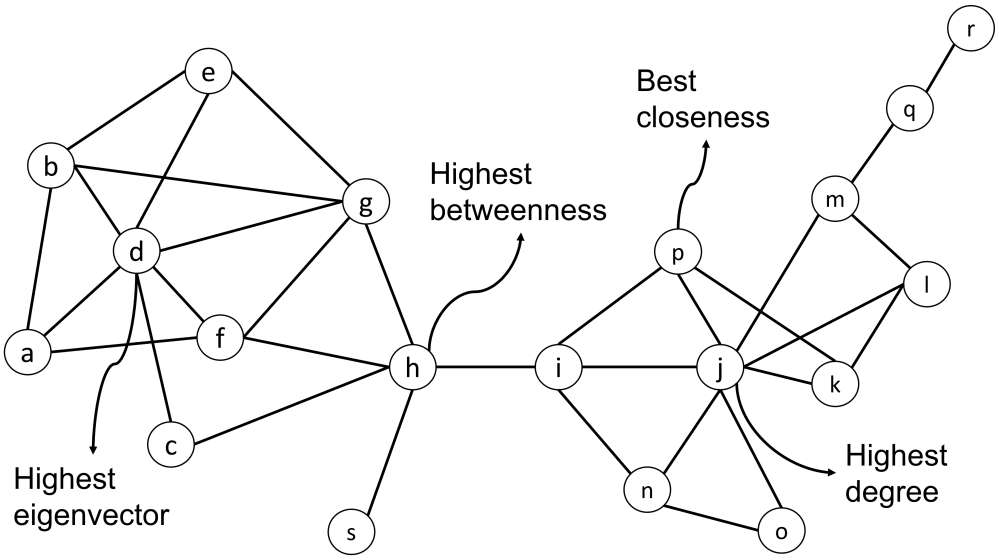


Figure 2.3: Example of different centrality measures

the network is. The network density D for an undirected graph without self-loops is:

$$D = \frac{2m}{n(n-1)},$$

where $m = |E|$ is the total number of edges and $n = |V|$ is the total number of vertices in the network. The range of D is between 0 and 1, where 0 indicates a completely disconnected graph and 1 indicates a fully connected network with every possible edge.

- **Centrality Measures.** Centrality measures quantify the importance of a vertex in the network. These measures help identify the key actors that play critical roles in information flow, influence, or connectivity. The notion of importance varies depending on the network and the analysis goals, so several different centrality measures have been proposed. Some of the most widely used measures are illustrated in Figure 2.3 and described below.

  – **Degree Centrality [31].** Measures the importance of a vertex by counting

its number of connections. A vertex with the highest degree centrality is often considered influential or central. The equation for degree centrality for a vertex v is $C_D(v) = d_v$, where d_v is the degree of vertex v. In Figure 2.3, vertex j has the highest degree centrality.

– **Closeness Centrality [107].** Measures the inverse of the sum of shortest path lengths from a vertex to all others. A vertex with high closeness centrality can reach all others efficiently, making it important for fast information spread. The equation for closeness centrality $C_C(v)$ is:

$$C_C(v) = \frac{1}{\sum_{u \in V} d(v,u)},$$

where $d(v,u)$ is the shortest path distance between v and u. In Figure 2.3, vertex p has the highest closeness centrality.

– **Betweenness Centrality [30].** Measures the fraction of shortest paths between all pairs of vertices that pass through a given vertex. A vertex with high betweenness centrality acts as a bridge, facilitating information flow. The equation for betweenness centrality $C_B(v)$ is:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

where σ_{st} is the total number of shortest paths between s and t, and $\sigma_{st}(v)$ is the number of those paths that pass through v. In Figure 2.3, vertex h has the highest betweenness centrality.

– **Eigenvector Centrality [11].** Assigns importance to a vertex based on the importance of its neighbors. A vertex with high eigenvector centrality is connected to other highly influential vertices, increasing its overall significance. The equation

for eigenvector centrality $C_E(v)$ is:

$$C_E(v) = \frac{1}{\lambda} \sum_{u \in N(v)} C_E(u),$$

where λ is the largest eigenvalue of the adjacency matrix and $N(v)$ is the set of neighbors of v. This definition creates a recursive relationship, since the centrality of v depends on the centralities of its neighbors. In practice, the centrality vector is obtained by solving the eigenvalue equation

$$\mathbf{A}\,\mathbf{C_E} = \lambda\,\mathbf{C_E},$$

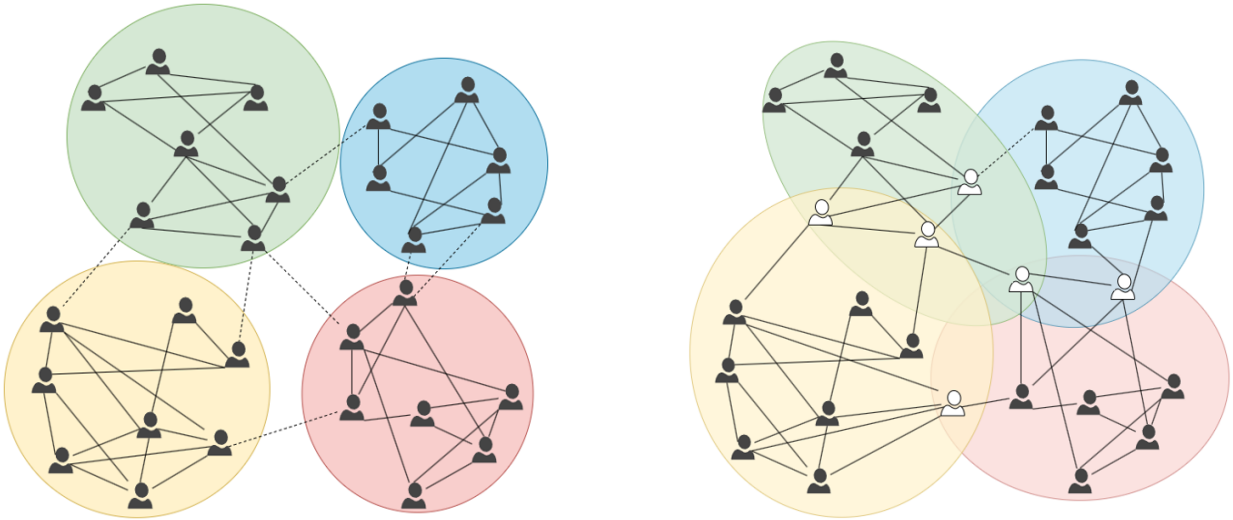
which yields a consistent set of centrality values corresponding to the dominant eigenvector of $\mathbf{A}$. In Figure 2.3, vertex d has the highest eigenvector centrality.

2.1.2 Communities in Social Networks

Identifying groups of entities that are similar or share common interests in a complex network is a fundamental task in SNA, commonly known as community detection. The concept of a community is somewhat subjective, and there is no universally accepted definition in the literature. A widely used definition views a community as a dense subgraph, where vertices within the community are more strongly connected to one another than to the rest of the network.

From a graph-theoretic perspective, given an undirected network $\mathcal{G} = (V, E)$, a community C is a subgraph $C = (V_c, E_c)$ with $V_c \subseteq V$ and $E_c \subseteq E$, such that the density of edges among vertices in V_c is higher than the density of edges between V_c and $V \setminus V_c$. Community detection aims to find a set of communities $C = \{C_1, C_2, \ldots, C_k\}$ such that $V = \bigcup_{i=1}^{k} V_{C_i}$ [82].

Different types of community structures exist, shaped by entities' characteristics and



(a) Disjoint community (b) Overlapping community

Figure 2.4: Examples of disjoint and overlapping community structure

diverse relationships. Some of the most common structures are shown in Figure 2.4 and Figure 2.5.

- **Disjoint Communities [84].** These are communities in which each vertex belongs to exactly one group, forming a partition or cluster of the network. In this context, a cluster refers to a subset of vertices that are grouped together based on structural similarity or connectivity patterns. A set of communities is disjoint when these clusters do not share any vertices, meaning every vertex is assigned to one and only one community. Figure 2.4a illustrates an example in which the network is divided into non-overlapping clusters.

- **Overlapping Communities [72].** Communities where a vertex may belong to multiple groups simultaneously, reflecting the fact that entities in real-world networks often participate in several relationships at once. Figure 2.4b shows vertices that are members of multiple overlapping groups. For example, an individual may

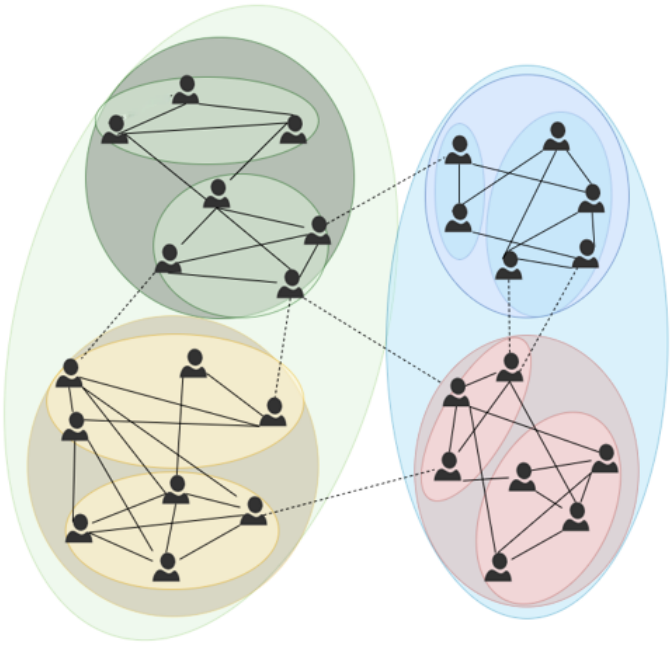


Figure 2.5: Hierarchical community structure

simultaneously be part of "Project Team 1," "Project Team 2," and a "Departmental Team" within a professional network.

- **Hierarchical Communities [9].** Communities organized in a nested, tree-like fashion, where large communities consist of smaller sub-communities. In an academic collaboration network, this could correspond to "Research Groups" at a high level and "Project Teams" within each research group. Figure 2.5 illustrates a hierarchical community structure.

Understanding these community structures is crucial for real-world network analysis. The choice of which community definition to use often depends on the characteristics of the network and the goals of the research. In the next section, we review community detection approaches designed to capture these structures.

2.1.3 Community Detection

Community detection aims to uncover the modular structure of a network by grouping vertices that are more densely connected to one another than to the rest of the network. Numerous approaches have been proposed, including graph partitioning, modularity optimization, hierarchical clustering, and spectral methods [49]. The choice of method depends on the network characteristics and the analysis objectives.

In this dissertation, we focus primarily on modularity-based and spectral approaches for detecting communities. The Louvain algorithm, a greedy modularity optimization method, and spectral clustering, based on the eigen-decomposition of the graph Laplacian, are used as our main techniques for community detection. Detailed descriptions of these algorithms are provided in Chapter 4, where they are applied in our methodology.

2.1.4 Community Evaluation Metrics

Detecting the suitable community structure in a large or complex network is an NP-Hard problem, as proven under various formulations and constraints [28, 52, 54]. Assessing the quality of discovered communities is further complicated in real-world networks where ground-truth community labels are often unavailable. Consequently, researchers have proposed various metrics to evaluate the quality of detected communities both in the presence and absence of ground-truth labels.

Ground-Truth Based Metrics When a ground-truth partition is available, evaluation metrics compare the detected communities with the true communities. Let $\mathcal{G} = (V, E)$ be a network with detected community set $C = \{c_1, c_2, \ldots, c_k\}$ and ground-truth community set $C' = \{c'_1, c'_2, \ldots, c'_{k'}\}$. Some commonly used metrics are:

- **Adjusted Rand Index (ARI) [44].** ARI measures the agreement between two partitions, adjusted for chance. For any similarity measure H, its chance-adjusted

version is:

$$H_c = \frac{H - \mathbb{E}(H)}{H_{\max} - \mathbb{E}(H)},$$

where $H_{\max}$ is the maximum value of H and $\mathbb{E}(H)$ is its expected value under a random model. The ARI is then:

$$ARI(C, C') = \frac{\sum_{ij}\left[\binom{N_{c_i,c'_j}}{2} - \mathbb{E}\binom{N_{c_i,c'_j}}{2}\right]}{\frac{1}{2}\left(\sum_i \binom{N_{c_i}}{2} + \sum_j \binom{N_{c'_j}}{2}\right) - \mathbb{E}\binom{N_{c_i,c'_j}}{2}},$$

where N_{c_i,c'_j} is the number of vertices shared between c_i and c'_j. ARI ranges from -1 (completely dissimilar) to $+1$ (perfect match).

- **Normalized Mutual Information (NMI) [117].** NMI measures the mutual information between two partitions:

$$NMI(C, C') = \frac{I(C, C')}{\sqrt{H(C)H(C')}},$$

where

$$I(C, C') = \sum_i \sum_j \frac{|c_i \cap c'_j|}{N} \log\left(\frac{N \times |c_i \cap c'_j|}{|c_i||c'_j|}\right), \qquad H(C) = -\sum_i \frac{|c_i|}{N} \log\left(\frac{|c_i|}{N}\right),$$

and $N = |V|$. NMI ranges from 0 (no mutual information) to 1 (perfect agreement).

- **Jaccard Index (JI) [38].** JI compares the similarity between two sets of vertex pairs:

$$JI(C, C') = \frac{|S \cap S'|}{|S \cup S'|},$$

where S is the set of vertex pairs that appear in the same community in partition C, and S' is the corresponding set of co-assigned vertex pairs in partition C'. JI ranges from 0 to 1, with 1 indicating identical partitions.

Unsupervised Metrics When no ground-truth communities are available, intrinsic metrics are used to assess the quality of a partition based on network structure alone.

- **Modularity [83].** The most widely used intrinsic metric, modularity compares the density of intra-community edges with that expected under a random null model. For an undirected graph, modularity Q is defined as:

$$Q = \frac{1}{2m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\mathbb{A}_{\text{נב}} - \frac{d_i d_j}{2m} \right] \delta(c_i, c_j),$$

  where $\mathbb{A}_{\text{נב}}$ is the adjacency matrix, d_i is the degree of vertex i, $m = |E|$ is the number of edges, and $\delta(c_i, c_j) = 1$ if i and j are in the same community, 0 otherwise. Q ranges from -1 to $+1$, with higher values indicating stronger community structure. Modularity has a well-known tendency to overlook small communities and merge them into larger ones, a phenomenon referred to as the resolution limit [29], which reduces sensitivity to fine-grained structures.

- **Conductance [3].** Conductance measures the fraction of edges leaving a community relative to its total edge volume:

$$\text{Conductance}(C) = \frac{1}{|C|} \sum_{i \in C} \frac{\text{cut}(c_i, \bar{c_i})}{\min\{\text{vol}(c_i), \text{vol}(\bar{c_i})\}},$$

  where $\text{cut}(c_i, \bar{c_i})$ counts edges between c_i and the rest of the graph, and $\text{vol}(c_i)$ is the sum of degrees of vertices in c_i. Lower conductance values indicate better-separated communities.

- **Silhouette Score [106].** The silhouette score measures cohesion and separation for each vertex:

$$S(c_i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i)$ is the average distance from vertex i to other vertices in its own community, and $b(i)$ is the minimum average distance to vertices in any other community. Distances are typically shortest-path distances. Scores close to 1 indicate well-defined communities.

In this dissertation, we primarily use modularity as our unsupervised evaluation metric, given its wide adoption in community detection. Although modularity has a resolution limit, it provides a useful baseline for comparing different partitions and is well-suited for large-scale networks where ground-truth labels are unavailable.

2.2 Text Analytics for Content-Based Network Analysis

The growth of online platforms such as X (formerly Twitter), Facebook, blogs, and discussion forums has led to an abundance of digital content that reflects human interaction and collective behavior. Content-based data can take many forms—text, images, videos—but in this dissertation we focus on textual data. Integrating textual information into social network analysis (SNA) enhances our ability to connect entities by shared content, improve community detection, and enable more personalized recommendations.

For example, in a research collaboration network, vertices represent researchers and edges represent collaborations. By incorporating textual content such as publications and proposals, we can connect researchers with similar research interests, detect topical research communities, and uncover emerging research trends.

2.2.1 Text Preprocessing

Text preprocessing is a crucial step to clean and standardize text prior to analysis. The common steps include:

- **Cleaning:** Remove HTML tags, special characters, and convert to lowercase to ensure consistent text.

- **Tokenization:** Break down text into individual words or phrases (tokens). For example, the sentence "The quick brown fox" becomes "[the, quick, brown, fox]" after tokenization.

- **Stop-word Removal:** Remove high-frequency uninformative words such as "the," "and," "is," which helps reduce noise and dimensionality.

- **Stemming/Lemmatization:** Reduce words to their root or dictionary form. For example, stemming would convert "running" and "runs" to "run," while lemmatization converts "ate" to its dictionary form "eat."

Other optional steps include digit removal, named entity recognition, and part-of-speech tagging, depending on the analysis task [51].

2.2.2 Text Representation

After preprocessing, text is transformed into a numerical representation suitable for machine learning models. Several common approaches include:

- **Bag-of-Words (BoW) [39].** Represents a document as a vector of word counts over a vocabulary V. For a document d, let $f_{i,d}$ denote the frequency of term i in d. The resulting document vector is:

$$\mathbf{x}^{(d)} = [f_{1,d}, f_{2,d}, \ldots, f_{|V|,d}],$$

  where each entry counts how many times word i appears in d. BoW is simple and interpretable, but it ignores word order and context.

- **Term Frequency–Inverse Document Frequency (TF–IDF) [115].** Adjusts BoW by weighting words based on both their within-document frequency and their

rarity across the corpus. For a word i in document d:

$$\mathrm{tf}(i,d) = \frac{f_{i,d}}{\sum_{i' \in d} f_{i',d}}, \qquad \mathrm{idf}(i,D) = \log\left(\frac{N}{|\{d \in D : i \in d\}|}\right),$$

where $N = |D|$ is the number of documents in the corpus. The combined TF–IDF weight is:

$$\text{tf–idf}(i,d,D) = \mathrm{tf}(i,d) \times \mathrm{idf}(i,D).$$

This representation highlights words that are frequent in a document but rare in the corpus, thus more informative.

- **Sentence Embeddings.** Modern language models such as SciBERT [7] encode entire documents into dense vectors $\mathbf{h}_d \in \mathbb{R}^k$ that capture semantic similarity. Unlike BoW or TF–IDF, embeddings capture contextual relationships between words rather than relying solely on frequency counts, which can provide richer representations for similarity-based tasks. However, this dense representation comes at the cost of interpretability, since individual dimensions of $\mathbf{h}_d$ do not correspond to human-readable features such as specific words or topics.

In this dissertation, we use BoW (with topic modeling), TF–IDF, and SciBERT-based sentence embeddings for document representation, depending on the experiment and baseline comparison.

2.2.3 Topic Modeling

Topic modeling discovers latent themes in a document collection by grouping documents that share similar content. Any topic modeling method that yields document–topic probability distributions can be used to represent documents in a topic space. Among the widely used methods, Latent Dirichlet Allocation (LDA) [8] provides a probabilistic

framework that models each document as a mixture of topics, while modern approaches such as BERTopic [35] leverage transformer-based embeddings with clustering to generate topics from contextualized text representations.

In this dissertation, we use both LDA (as a probabilistic topic model) and BERTopic (for embedding-based topic modeling). The detailed generative process of LDA and the BERTopic modeling pipeline are presented in later chapters.

2.2.4 Topic Coherence Measures

Topic models require specifying the number of topics beforehand, and since they are unsupervised, there is no direct ground-truth measure to evaluate topic quality. Topic coherence measures are widely used to assess the interpretability of topics and to select an appropriate number of topics. The key intuition is that a good topic should consist of words that are semantically related and co-occur frequently in similar contexts.

- **`C_V` Coherence [101].** The `C_V` measure combines a sliding-window co-occurrence model, normalized pointwise mutual information (NPMI), and cosine similarity between context vectors. For a topic t with top M words $W = \{w_1, w_2, \ldots, w_M\}$, let $\mathbf{v}(w_i)$ denote the NPMI-based context vector for word w_i, and let the topic vector be the average of all context vectors:

$$\mathbf{v}_{\text{topic}} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{v}(w_i).$$

  The `C_V` coherence for topic t is then given by:

$$C_V(t) = \frac{1}{M} \sum_{i=1}^{M} \cos\big(\mathbf{v}(w_i), \mathbf{v}_{\text{topic}}\big), \qquad \cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}.$$

Each component of the context vector $\mathbf{v}(w_i)[j]$ is defined using normalized PMI:

$$\mathbf{v}(w_i)[j] = \text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)},$$

where $P(w_i, w_j)$ is the probability of observing w_i and w_j in the same sliding window and $P(w_i), P(w_j)$ are their marginal probabilities. Higher C_V values indicate more coherent and interpretable topics.

- **UMass Coherence [79].** The UMass measure is based on document co-occurrence counts from the original corpus. For topic t with M top words:

$$C_{\text{UMass}}(t) = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log\left(\frac{D(w_i, w_j) + 1}{D(w_j)}\right),$$

where $D(w_i, w_j)$ is the number of documents containing both w_i and w_j, and $D(w_j)$ is the number of documents containing w_j. A higher score indicates that the top words of the topic co-occur more frequently in the corpus, suggesting a more coherent topic.

Other coherence measures such as UCI and NPMI also exist. Following the findings of Röder et al. [101] that `C_V` correlates strongly with human interpretability, we use `C_V` coherence to select the optimal number of topics and also report UMass coherence as an additional point of comparison.

2.3 Summary

In summary, this chapter outlined the theoretical foundations and analytical techniques that inform our subsequent methodology. We reviewed core concepts of social network analysis, including structural properties, community definitions, and evaluation metrics, followed by text analytics methods essential for integrating content-based information into

network models. Together, these components provide the conceptual and computational basis for the topic-based community detection and recommendation framework developed in the following chapters.

CHAPTER THREE

DATASET

One of the primary goals of this dissertation is to leverage content-based resources to construct scholarly social networks and generate collaboration recommendations based on the constructed networks. To achieve this, we focus on scholarly data as the foundation for validating our proposed approach of integrating content and social network information.

Our dataset is drawn from three institutions: Montana State University (MSU), Washington State University (WSU), and Colorado State University (CSU). We began by collecting data from our home institution, MSU, and later expanded the dataset to include WSU and CSU, as these institutions are considered MSU's closest peer universities in terms of research profile and disciplinary breadth. This expansion allows us to evaluate the generalizability of our methods across multiple but comparable academic settings.

The final dataset comprises over $1,600$ researchers and more than $45,000$ publications across the three institutions, spanning the years 2004 to May 2025. Each record contains metadata such as faculty name, department, academic rank, and associated publication titles and abstracts. This collection represents a diverse set of disciplines, ranging from engineering and natural sciences to social sciences and humanities, making it well-suited for studying interdisciplinary collaboration opportunities.

In this chapter, we present (i) a detailed description of the data sources and collection process, (ii) the cleaning and preprocessing steps applied to ensure high-quality and consistent textual and network data, and (iii) descriptive statistics that characterize the dataset. These statistics provide insight into the distribution of researchers, publications, and research areas across the three institutions. The processed dataset described here serves as the foundation for all subsequent experiments and analyses presented in the remainder of

this dissertation.

3.1 Data Source

3.1.1 Institutional Data

To build our dataset, we collected current faculty information from three institutions: MSU, WSU, and CSU.

For MSU, our home institution, we were able to obtain faculty data directly through the university library in May 2025. This dataset includes a total of 606 faculty records with fields such as first and last name, academic rank, college, department, and email address. Based on this data, we identified approximately 7 major colleges and around 48 academic departments, giving a reasonable representation of MSU's disciplinary structure.

For WSU, no internal dataset was available, so we utilized the university's public academic catalog, available at `https://catalog.wsu.edu/`. The catalog was collected in May 2025, and we restricted our data collection to the Pullman campus (the main WSU campus), excluding other branches such as Tri-Cities and Vancouver to maintain consistency. We parsed the catalog to extract current faculty records, including first and last name, academic rank, and college affiliation. This process yielded $1,816$ faculty entries across approximately 11 major colleges.

For CSU, we relied on the public HTML-based faculty directory, available at `https://catalog.colostate.edu/general-catalog/faculty/`, which we collected in May 2025. The directory provided first and last name and academic rank for each faculty member. Unfortunately, the CSU directory did not contain college or department affiliation. Although having this information would allow a clearer view of interdisciplinary collaboration potential, the CSU data still provides a valuable opportunity to analyze collaboration patterns and research topics using textual content. This dataset contains $1,942$ faculty entries.

Across the three institutions, these datasets collectively provide a broad and diverse set of researchers, enabling the construction of content-based networks that span multiple disciplines and institutional contexts.

3.1.2 Publication Data

In this work, we primarily rely on publication metadata (i.e., titles and abstracts) from researchers to power our content-based methods. The intuition behind using publication content is that titles and abstracts can reveal latent relationships between researchers that go beyond prior collaborations or citation links. By learning topic models to identify topical similarity, we can capture both existing and potential collaborations, thus supporting the discovery of interdisciplinary opportunities.

Our primary source of publication metadata is "OpenAlex" [93], a modern scholarly catalog developed as a successor to the Microsoft Academic Graph (MAG). OpenAlex aggregates data from multiple sources, such as CrossRef, PubMed, and publisher feeds, to provide a comprehensive view of scholarly works. Using the OpenAlex API endpoint `https://api.openalex.org/works`, we collected all works associated with MSU, WSU, and CSU. Each institution is identified in OpenAlex with a unique institutional identifier: MSU — `I23732399`, WSU — `I72951846`, and CSU — `I92446798`. Our data collection spanned works published from `2004-01-01` to `2025-05-20`, allowing us to focus on publications where metadata (such as abstracts and affiliations) are more complete. This process yielded a total of $22,150$ works for MSU, $72,961$ works for WSU (Pullman campus only), and $70,390$ works for CSU.

We implemented the data collection using the Python library Diophila [1], which handles API paging and filtering. The returned records, referred to as "works" in OpenAlex, are structured as nested dictionaries containing detailed metadata fields. These include a unique

[1] `https://github.com/smierz/diophila`

OpenAlex work ID, title, abstract (when available), publication date, venue, publication type (article, book, or conference paper), DOI, open access status, and a list of authors with their affiliations and ORCID identifiers (if present). Given the relational nature of the data, we stored the results in separate MySQL databases for each institution, enabling efficient filtering and matching with faculty rosters in later steps.

Although OpenAlex provides the ability to query works at the individual author level, initial experiments at MSU revealed issues with name ambiguity, leading to misattributed works even when institutional filters were applied. To mitigate this, we chose to collect works at the institutional level and then match them post-hoc to our cleaned faculty lists by name. This conservative approach may exclude some legitimate works but ensures that the publications we retain are reliably attributed to the correct faculty.

Finally, while OpenAlex returns a wide range of fields, for the purposes of this dissertation we primarily focus on the "title" and "abstract" fields as the textual basis for constructing the content-based scholarly networks. Additional metadata such as date, type, and venue are retained for filtering and descriptive statistics but are not directly used in the network construction process.

3.2 Data Cleaning and Preprocessing

3.2.1 Institutional and Publication Record

Preparing the raw institutional and publication data for analysis is a critical step to ensure reliable alignment between faculty and their scholarly output. The goal of this stage is to transform heterogeneous and noisy records into a consistent dataset that can support our content-based modeling. Without careful cleaning, issues such as duplicate names, mismatched affiliations, and incomplete abstracts would undermine the quality of the networks we construct.

The MSU catalog obtained from our institutional library contained entries for both

faculty and staff. To focus only on current research-active faculty, we filtered based on academic rank, retaining positions such as professor, associate professor, assistant professor, and research professor. For WSU and CSU, the catalogs did not include explicit academic rank codes, so we initially retained all entries. This means some individuals may not have active publication profiles in OpenAlex, but they were removed later when we align faculty names with actual publications.

Duplicate names presented another issue. At MSU, we identified three researchers with identical names. After manual inspection, we found that only one of them had a valid OpenAlex profile, while the other two had minimal research output (fewer than five publications). To remain consistent across institutions, we excluded those two entries. At WSU and CSU, no duplicate faculty names were found. Ideally, such alignment would be done with unique identifiers such as ORCID, but since this information was not available in the catalogs, we relied on name matching.

The next step was to match institutional publication records with the faculty rosters. We had first and last names for all three institutions, but matching these to OpenAlex author records is challenging because publication names often differ slightly—for example, the presence or absence of middle initials. To address this, we used wildcard queries in MySQL (e.g., "first%last") when matching against OpenAlex's author tables. Fortunately, OpenAlex includes an `alternate_display_name` field, which records common name variations, and a `last_known_institution` field, which indicates the author's current affiliation. By requiring both a name match and a matching institution, we reduced the risk of false positives. This approach may omit some legitimate works, but it ensures that the publications we retain are reliably attributed to the correct faculty.

After aligning researchers with their works, we cleaned the publication records, focusing on the title and abstract fields. Although OpenAlex provides metadata for each work, abstracts are sometimes missing (especially in books) or incomplete—for example,

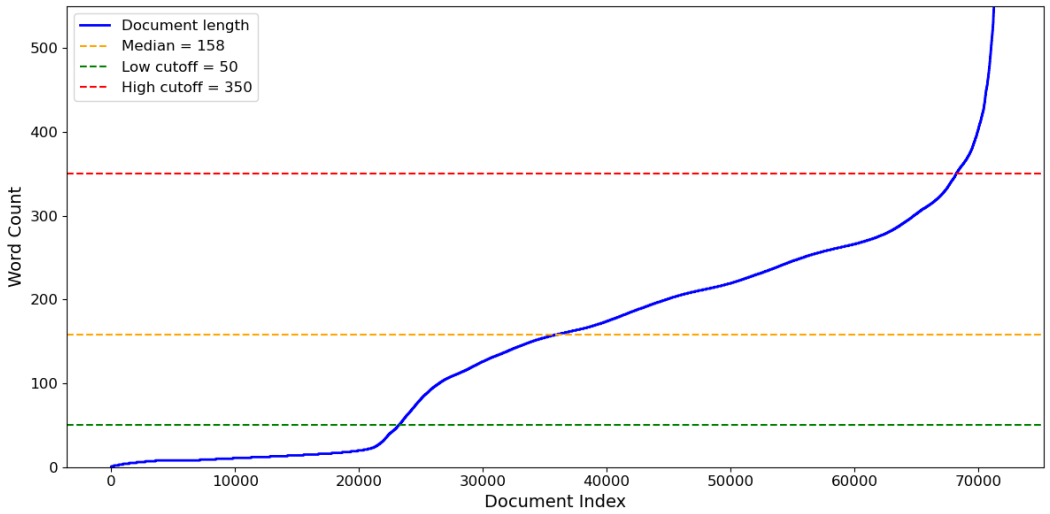


Figure 3.1: Distribution of word counts per document before cleaning

truncated with "...". To improve reliability for content-based analysis, we applied word-count thresholds. Publications with fewer words than the threshold were excluded because such short documents provide too little information to extract meaningful topics, and topic modeling on very sparse text often fails to capture coherent themes. On the other hand, some records contained extremely long text segments (often over 1,000 words), which in practice represent full-text dumps rather than true abstracts. Including these overly long entries risks introducing noise and distorting topical distributions. By setting both lower and upper cutoffs, we retained abstracts of sufficient length to capture meaningful topical signals while filtering out entries that were either too short or too long to be useful.

Following this step, we also excluded researchers with fewer than five valid publications. Such low-volume profiles do not provide enough textual evidence for stable topic extraction, and including them would likely result in noisy or unrepresentative topical assignments. This additional filter ensured that each retained researcher contributed a sufficient body of work to support meaningful content-based analysis.

Figure 3.1 shows the distribution of title and abstract word counts across all three

Table 3.1: Faculty and publication counts across processing stages

Institution	**Faculty Count**			**Publication Count**		
	Raw	**Matched**	**Cleaned**	**Raw**	**Matched**	**Cleaned**
MSU	606	404	276	22,150	10,062	6,372
WSU	1,816	1,004	613	72,961	27,037	15,931
CSU	1,942	1,105	745	70,390	34,848	23,072
Total	4,364	2,513	1,634	165,501	71,947	45,375

institutions along with the chosen cutoffs. Table 3.1 summarizes the corresponding filtering steps. The "raw" stage reflects the full faculty rosters provided by the institutions together with the complete set of publications retrieved from OpenAlex using institutional identifiers. The "matched" stage reports the subset of faculty for whom OpenAlex author records could be reliably identified through name and affiliation matching, along with the publications attributed to those authors. The "cleaned" stage applies the final content-based filters, retaining only faculty with at least five matched publications and limiting documents to those with title and abstract lengths within the chosen word-count thresholds. These steps collectively yield the final dataset used in this dissertation.

3.2.2 Text Preprocessing

After cleaning the faculty and publication records, we applied text preprocessing to the titles and abstracts to prepare them for later stages of analysis. The purpose of this step was to normalize text, remove noise, and retain words that carry meaningful semantic information. To ensure each publication was treated as a coherent unit, we combined the title and abstract of every work into a single document before applying preprocessing.

The motivation for preprocessing is twofold. First, traditional topic models such as

LDA, rely on bag-of-words representations, and without normalization and filtering, they are easily dominated by noise and spurious tokens. In contrast, modern embedding-based models such as BERTopic are less sensitive to preprocessing when it comes to discovering topics, since embeddings can capture contextual meaning directly from raw text. However, consistent preprocessing still plays a critical role in tasks such as word cloud visualizations, where noisy tokens, symbols, or short fragments would otherwise obscure the main topical themes. In this sense, preprocessing ensures that researcher-level visualizations remain cleaner and more interpretable.

Our pipeline began by lowercasing all text and normalizing common symbols (e.g., middle dots converted to periods, various dash characters standardized to hyphens). We then removed numerical and symbolic patterns that do not contribute to topical meaning, including dosage expressions (e.g., 5mg), numeric ranges (e.g., 123–456), temperature markers (e.g., 40°C), and equations of the form variable = number (e.g., $\beta = 0.5$). URLs, DOIs, and similar metadata were also stripped. These steps helped reduce vocabulary size and eliminated noise unlikely to aid interpretability.

Next, we applied tokenization and lemmatization using the `spaCy` library[2]. Lemmatization reduced inflected forms to their base form (e.g., studies → study), and stopwords were removed to retain only content-bearing terms. To further refine the text, we restricted tokens to alphabetic words and excluded very short tokens, requiring a minimum word length greater than two characters. The output was a cleaned, lemmatized, and normalized version of each document, represented as a space-separated sequence of tokens.

By enforcing these preprocessing steps, we aimed to balance two goals: ensuring compatibility with topic modeling techniques and producing clearer, noise-free visualizations of topical content. This alignment ultimately helps the constructed content-based networks more accurately reflect meaningful areas of expertise.

[2] `https://spacy.io/`

Table 3.2: Summary statistics of researchers in the final dataset

Institution	**#Researchers**	**#Publications**	**Mean**	**Median**	**Max**	**Std**
MSU	276	6,372	27.86	18	168	27.13
WSU	613	15,931	31.05	18	237	34.53
CSU	745	23,072	37.75	25	246	39.42

3.3 Data Statistics

To better understand the characteristics of the cleaned dataset, we begin by examining the distribution of researchers and their publication records across the three institutions. Although some of the original data were collected from institutional faculty catalogs, after applying publication-based filters we refer to the remaining set as "researchers", since they represent the active research population used in our analysis. Table 3.2 provides an overview of the dataset, including the number of researchers and publications per institution, along with descriptive statistics of publication counts per researcher.

As shown in Table 3.2, MSU contributed the smallest number of researchers and publications overall, while WSU and CSU account for substantially larger totals. Nevertheless, the median number of publications per researcher is broadly comparable across institutions, with both MSU and WSU reporting a median of 18 and CSU slightly higher at 25. The higher means relative to the medians indicate that publication counts are not evenly distributed, but are instead skewed by a subset of highly prolific researchers. This skew is most evident at CSU, where the maximum researcher publication count reaches 246, compared to 237 at WSU and 168 at MSU.

To further illustrate this variation, Figure 3.2 presents violin plots of the publication distributions per researcher. All three institutions display heavily right-skewed distributions:

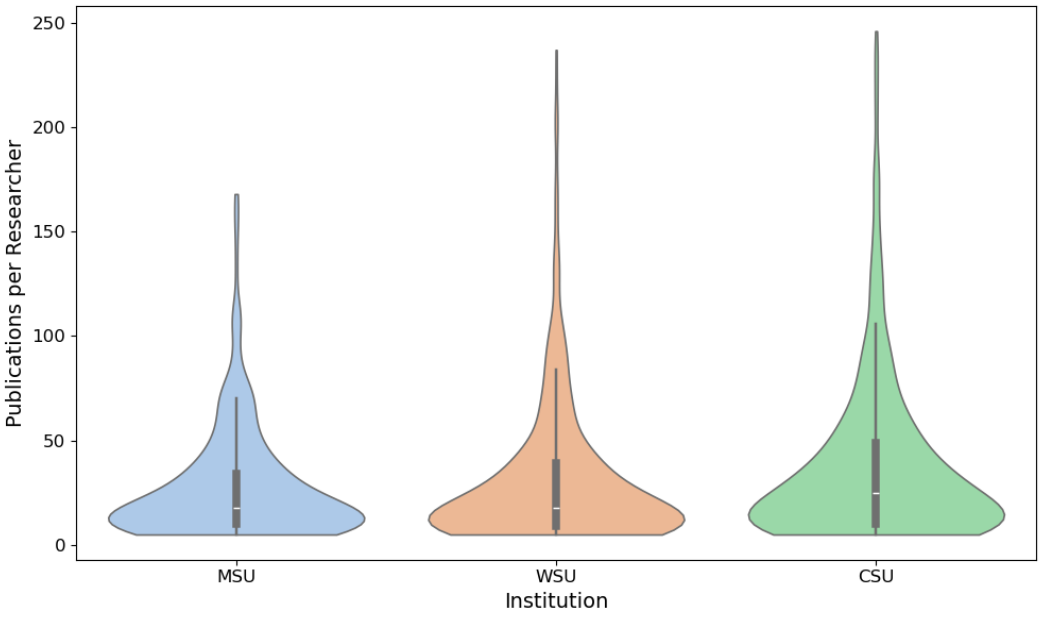


Figure 3.2: Distribution of publication counts per researcher across institutions

the majority of researchers have modest publication counts, while a smaller fraction exhibit much higher outputs. We highlight these imbalances here for completeness, but a more detailed examination of their implications is reserved for Chapter 6.

Figure 3.3 shows the number of publications per year for each institution between 2004 and 2025. All three institutions demonstrate overall upward growth in research output, with publication activity peaking around 2020–2021. The slight dip in the final year is expected, as the 2025 data only cover publications indexed up to May. Although the data are current to that point, some publications may still not appear due to natural delays in the publishing process as well as the time it takes for OpenAlex and its sources to incorporate newly released works. While these temporal patterns help validate dataset coverage, temporal dynamics are not the main focus of this dissertation and are not examined further.

Table 3.3 summarizes coauthorship patterns both within and across institutions. The diagonal entries represent works with two or more authors from the same institution,

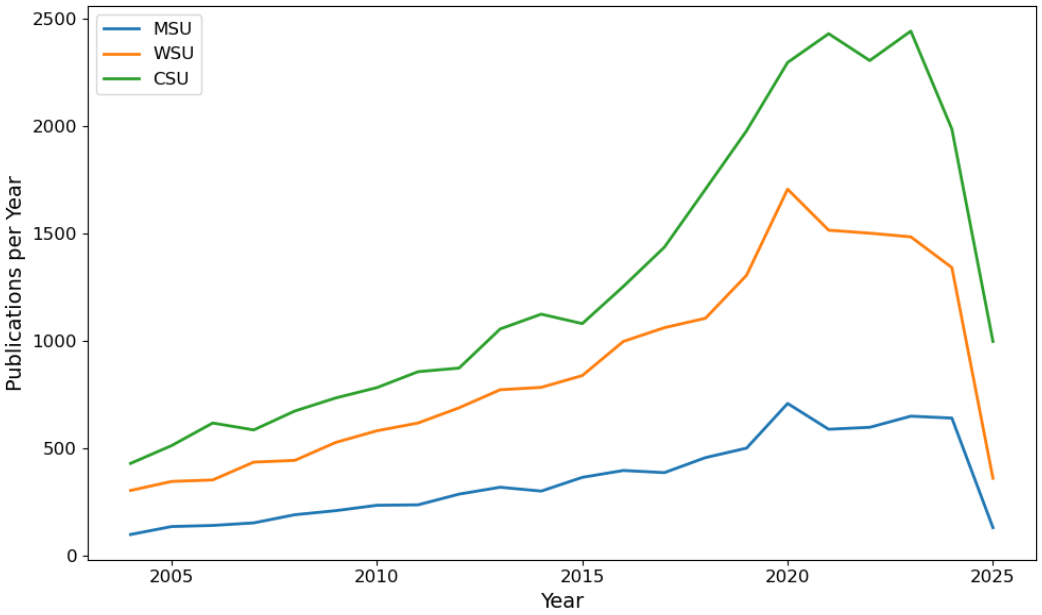


Figure 3.3: Annual publication counts per institution

Table 3.3: Publications coauthored within and across institutions

Institution	**MSU**	**WSU**	**CSU**
MSU	1,014	40	23
WSU	40	2,558	116
CSU	23	116	3,932

while the off-diagonal entries capture cross-institutional collaborations. As expected, most collaborations occur within institutions, but a smaller number span multiple institutions, providing useful context for the study of scholarly networks developed in later chapters.

3.4 Summary

In summary, this chapter detailed the institutional and publication data used throughout this dissertation, along with the cleaning, preprocessing, and descriptive statistics that

shaped the final dataset. By integrating faculty catalogs with publication metadata from OpenAlex, we constructed a research-ready dataset capturing over 1,600 active researchers and more than 45,000 publications across MSU, WSU, and CSU. The descriptive analyses highlighted variation in researcher productivity, temporal growth in publication output, and both intra- and cross-institutional collaboration patterns. Together, these characteristics provide a foundation for the methodological developments and analyses presented in the subsequent chapters.

CHAPTER FOUR

COMMUNITY DETECTION FOR SCHOLARLY COLLABORATION

In this chapter, we address the first research question of this dissertation: "How can topic modeling be integrated with social network analysis to promote interdisciplinary collaboration recommendations?" We explore the limitations of traditional recommender approaches in scholarly networks, particularly their tendency to reinforce existing disciplinary boundaries. To overcome these challenges, we propose constructing a scholarly network based on topic similarity derived from publication metadata and applying community detection methods to reveal both disciplinary and interdisciplinary clusters. The chapter begins with the motivation and problem statement, followed by a review of related works. We then present the methodology, including topic modeling, network construction, and community detection, before discussing the results and concluding with a summary.

4.1 Motivation and Problem Statement

Traditional recommender systems based on collaborative filtering rely on users' past interactions to generate new suggestions. While these methods are effective in many domains, they often face the issue of overspecialization [1]. This problem occurs when a system continues to recommend items that are too similar to what the user has already consumed, limiting the diversity of options. In the scholarly setting, this translates into recommending collaborators who are already part of the same disciplinary or co-authorship circles, rather than exposing researchers to new or interdisciplinary opportunities. Content-based methods are often better at introducing diversity in recommendations [102] because they focus on the content of the items themselves. In academic networks, this means going beyond historical collaboration data and using the actual topics of publications as the basis for

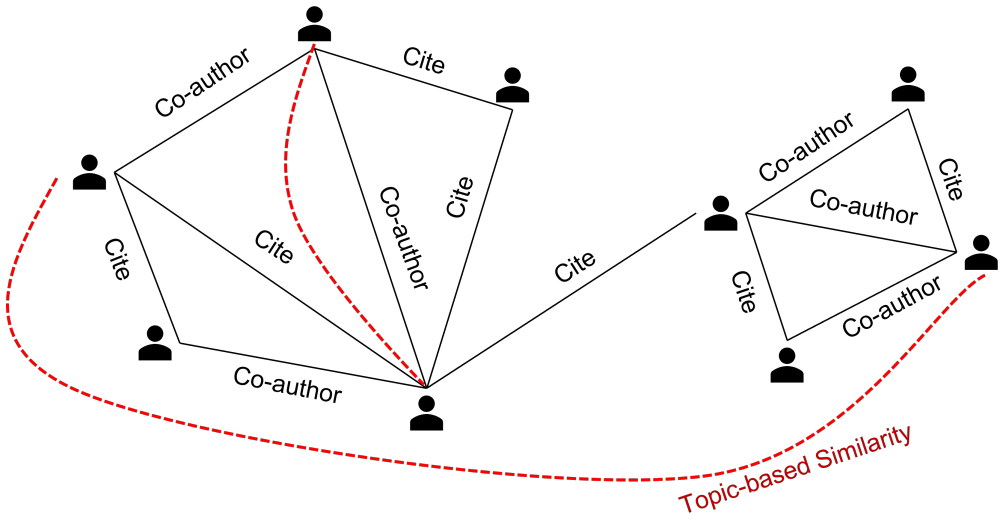


Figure 4.1: Sample scholarly network based on direct relationships

making connections.

Scholarly networks can be constructed in many different ways depending on the data source. For example, a co-authorship network connects two researchers if they have published together, while a citation network connects publications or authors based on citation links. Other variations include institutional networks, where affiliation data is used to form links, or even social platforms where researchers follow or interact with each other online. Each type of network answers different questions. A co-authorship network is useful for finding frequent collaborators and established research groups, while a citation network is better suited to identifying influential publications or authors and following research trends [41, 59]. These network types are valuable but they are often limited in terms of interdisciplinary discovery. A co-authorship network, for instance, will mostly recommend people who are already connected through direct collaborations, while a citation network may overemphasize disciplinary influence.

Figure 4.1 illustrates this limitation. The dense subgraphs correspond to disciplinary groups where researchers frequently co-author or cite each other. However, researchers

outside these groups may share topical interests similarities not captured by these direct connections. For example, a computer scientist developing machine learning models and an agricultural scientist applying machine learning to crop efficiency are unlikely to appear in the same co-authorship cluster, yet they share a common topical interest that could lead to meaningful interdisciplinary collaboration.

Problem Statement: Given a collection of researchers' publication metadata, specifically titles and abstracts, how can we construct a scholarly network based on topic similarity in order to detect communities that not only capture traditional disciplinary clusters but also uncover potential interdisciplinary collaborations?
Hypothesis: We hypothesize that a scholarly network constructed using topic-based relationships will reveal research communities grounded in similar topical interests, where members may come from different departments or disciplines. Such communities can identify interdisciplinary collaboration opportunities that would not be discovered through networks built only on direct relationships such as co-authorship or citation.

To investigate this, we derive topic distributions for each researcher using Latent Dirichlet Allocation (LDA). We calculate pairwise similarity between researchers with Jensen–Shannon divergence, which produces a fully connected network where each researcher is linked to every other. Since such a network is too dense for meaningful analysis, we apply edge weakening to emphasize stronger connections. Finally, we employ community detection algorithms such as Louvain and Spectral clustering to discover communities of researchers that can serve as the basis for cross-domain collaboration recommendations.

4.2 Related Work

Several approaches have been proposed to recommend research collaborators using scholarly data. Broadly, these can be grouped into three categories: methods based solely on co-authorship relations, content-based approaches that rely on textual or topical similarity, and hybrid methods that integrate multiple data sources. Below, we review representative works from each of these categories.

Early work often focused on homogeneous scholarly networks, particularly co-authorship relations. Nowell et al. [69] framed the link prediction problem, asking how new ties can be inferred from existing network snapshots. Using a co-authorship network, they proposed neighborhood- and path-based measures to predict future collaborations. Later, Backstrom et al. [4] introduced a supervised random walks algorithm to solve the link prediction problem in co-authorship networks, while Li et al. [63] incorporated academic metrics such as authorship order and collaboration frequency into random walk models. These approaches rely exclusively on topological features, and while useful for modeling future ties, they tend to reinforce disciplinary boundaries rather than reveal interdisciplinary opportunities.

Other works have turned to content-based data, especially textual information from academic publications, blogs, or microblogs. Velardi et al. [120] proposed a content-based method for network evolution, detecting relevant concepts and emerging themes through clustering of TF–IDF vectors. Liang et al. [67] proposed a time-aware topic recommendation model for microblogs, using cosine similarity between user topics and candidate topics while incorporating temporal dynamics. For research collaboration, Liang et al. [68] applied LDA topic modeling to discover potential research fields, then measured similarity between author-specific topic profiles to provide cross-disciplinary recommendations. Their approach focused primarily on topical similarity between individuals and did not model the broader structural relationships among researchers. In contrast, this dissertation integrates topic modeling

with social network analysis, using similarity-derived edges to construct a scholarly network whose community structure provides additional context for understanding research groups and potential collaborations.

Kong et al. [56] later introduced the Beneficial Collaborator Recommendation model, which incorporated both dynamic research interests and academic influence, weighting recent topics more heavily. These content-based approaches highlight the role of topical similarity in broadening collaboration opportunities, but most stop short of integrating network-level structure or community detection.

Hybrid methods combine heterogeneous sources such as co-authorship, expertise, and institutional data. Yang et al. [130] used a heterogeneous network incorporating expertise, co-authorship, and affiliation features, with semantic similarity calculated via language models and ranking performed with SVM-Rank. Kong et al. [55] employed Word2Vec on publication titles to generate concept vectors, combining them with co-authorship links and random walk methods to recommend collaborators. Zhou et al. [138] integrated multi-source data from ResearchGate (co-author, co-project, citations, following), and introduced a time-aware edge weighting strategy to capture the recency of collaborations. These hybrid approaches leverage richer data, but often at the cost of complexity and data availability.

In summary, co-authorship-based approaches excel at predicting future links but tend to reinforce disciplinary boundaries, while content-based methods highlight topical similarity yet often overlook structural features of networks. Hybrid models combine multiple data sources but may be difficult to generalize across institutions. Our approach is somewhat in between content-based and hybrid methods. We use the topical content of publications to build the network, but the analysis itself relies on social network techniques such as edge weakening and community detection. This allows us to capture both topical alignment and network structure without depending on direct ties like co-authorship or citation.

4.3 Methodology

This study builds on the dataset introduced in Chapter 3, which contains publication metadata (titles and abstracts) from researchers at MSU, WSU, and CSU. Accordingly, four datasets are considered: three institution-specific subsets and one combined dataset representing all institutions (denoted as MWC). These datasets serve as the foundation for topic modeling and subsequent network construction. This section describes the procedures used to derive latent topics, construct topic-based researcher networks, and apply community detection algorithms to identify potential interdisciplinary research communities.

4.3.1 Latent Topic Discovery

To construct a scholarly social network that reflects researchers' conceptual interests, the first step was to uncover the latent topics embedded within their publications. Each publication's title and abstract were combined to form a single document, and all research articles collected from OpenAlex constituted the full corpus for topic modeling.

Latent Dirichlet Allocation Latent Dirichlet Allocation (LDA), introduced by Blei et al. [8], is a generative probabilistic model designed to discover hidden thematic structures in large text corpora. The central idea is that documents can be represented as mixtures of latent topics, while each topic itself is a probability distribution over words. This framework allows researchers to infer abstract concepts from unstructured text without requiring labeled data, making LDA particularly suitable for scholarly corpora.

LDA assumes a fixed number of latent topics T and represents each document as a bag-of-words (ignoring word order). Documents are modeled as probability distributions over topics, governed by a Dirichlet prior α, while topics are modeled as probability distributions over words, governed by a Dirichlet prior η.

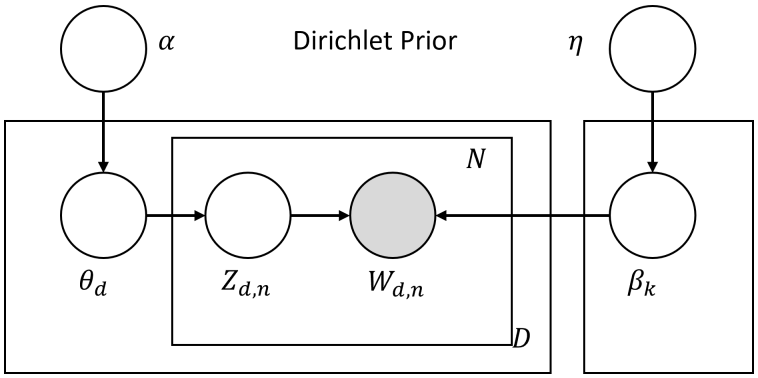


Figure 4.2: Plate diagram depicting the LDA generative process

Figure 4.2 shows the plate diagram for LDA's generative process. Here, D denotes the number of documents, N the number of words per document, θ_d the topic distribution for document d, and β_k the word distribution for topic k. The latent variable z_{dj} indicates the topic assignment for the j-th word in document d, while w_{dj} represents the observed word itself.

Formally, the generative process for T topics and word distributions $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_T\}$ proceeds as follows:

- For each topic $t = 1, \ldots, T$, draw a word distribution $\beta_t \sim Dirichlet(\eta)$
- For each document d, do:
  - Draw topic proportions $\boldsymbol{\theta}_d \sim Dirichlet(\alpha)$
  - For each word w_{dj} in document d, do:
    - Draw a topic assignment $z_{dj} \sim Multinomial(\theta_d)$
    - Draw a word $w_{dj} \sim Multinomial(\beta_{z_{dj}})$

In this framework, α and η are the hyperparameters for the Dirichlet priors, θ_d is the latent topic mixture for document d, and β_t specifies the probability distribution of words for topic t. Both θ_d and β_k are unobserved and must be learned.

Training an LDA model therefore involves estimating the latent distributions $\boldsymbol{\theta}_d$ and $\boldsymbol{\beta}_t$ given the observed corpus. Since exact inference is intractable, approximate methods such as Variational Bayes (VB) [50], Expectation-Maximization (EM) [21], or Gibbs sampling [33] are commonly used. Among these, Gibbs sampling, based on Markov Chain Monte Carlo, has proven effective and is the approach implemented in the version of LDA used in this study.

Implementation Detail In practice, titles and abstracts were concatenated into single documents and underwent a series of standardized preprocessing operations, including tokenization, stopword and punctuation removal, and lemmatization. A detailed description of these preprocessing procedures is provided in Section 3.2.2. Extremely frequent and rare terms, those appearing in more than 70% of documents or fewer than two times overall, were further excluded, resulting in vocabularies of size $9,093$ for MSU, $15,118$ for WSU, $18,468$ for CSU, and $27,061$ for the combined MWC dataset.

The topic modeling was performed using the MALLET implementation of LDA through the Gensim library [96], which employs an optimized Gibbs sampling procedure to estimate the LDA parameters. A key hyperparameter in LDA is the number of topics, which must be specified in advance. We evaluated all four datasets with multiple number of topics: for the institution-specific datasets (MSU, WSU, CSU), the number of topics ranged from 50 to 150 in increments of 10, and for the combined MWC dataset, the number of topics ranged from 100 to 200 in increments of 10, reflecting its larger size. The final number of topics for each dataset was chosen based on the highest `C_V` coherence score (Section 2.2.4).

Figure 4.3 shows the coherence scores for each dataset across the evaluated topic numbers. The blue marker indicates the maximum coherence score, which was selected as the best topic configuration for that dataset, corresponding to 80 topics for MSU and WSU, 90 topics for CSU, and 160 topics for the combined MWC dataset.

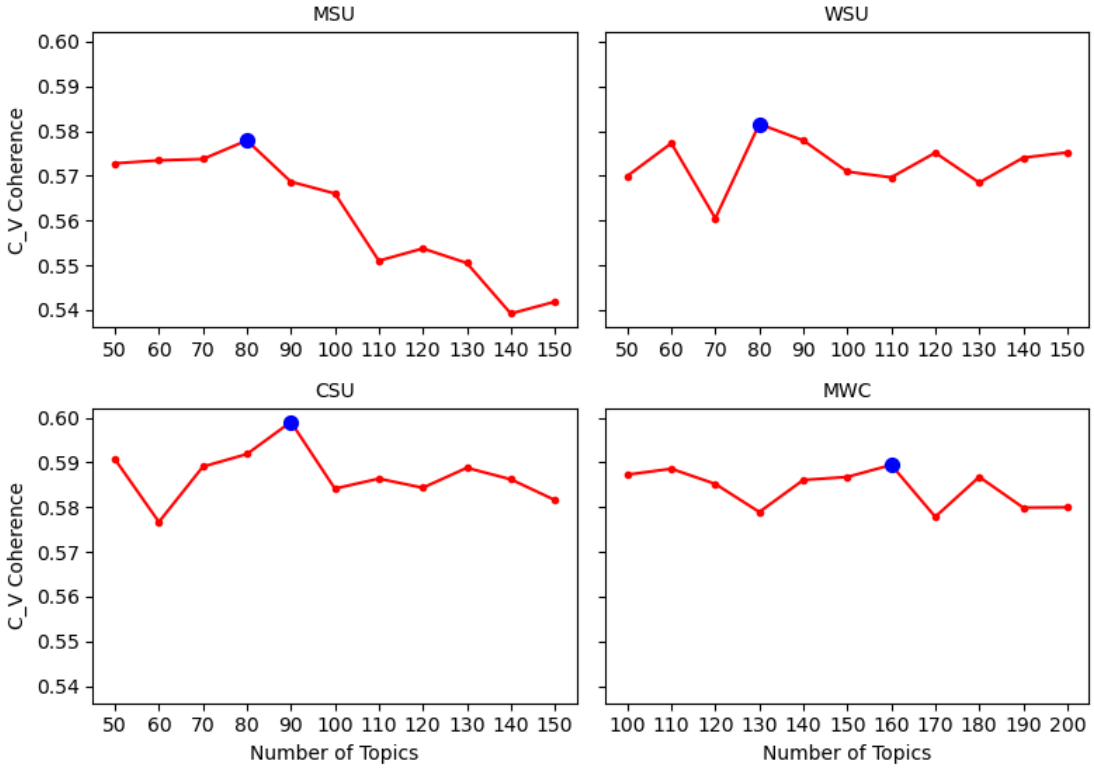


Figure 4.3: `C_V` coherence scores across topic numbers for the four datasets

To illustrate the quality and interpretability of the learned topics, Figure 4.4 presents four representative topics selected randomly from the LDA model trained on the combined MWC dataset. LDA produces a set of unsupervised topics in which every word in the vocabulary has a probability of appearing within each topic. For visualization, the 50 words with the highest probabilities are displayed as WordClouds, with larger words indicating higher probability within the topic.

Based on the prominent terms, Topic 86 highlights molecular and gene expression research, with terms such as transcription, RNA, and regulation. Topic 109 focuses on genetics and trait identification, including words like marker, genome, and population. Topic 143 centers on infectious disease research with terms such as virus, transmission,

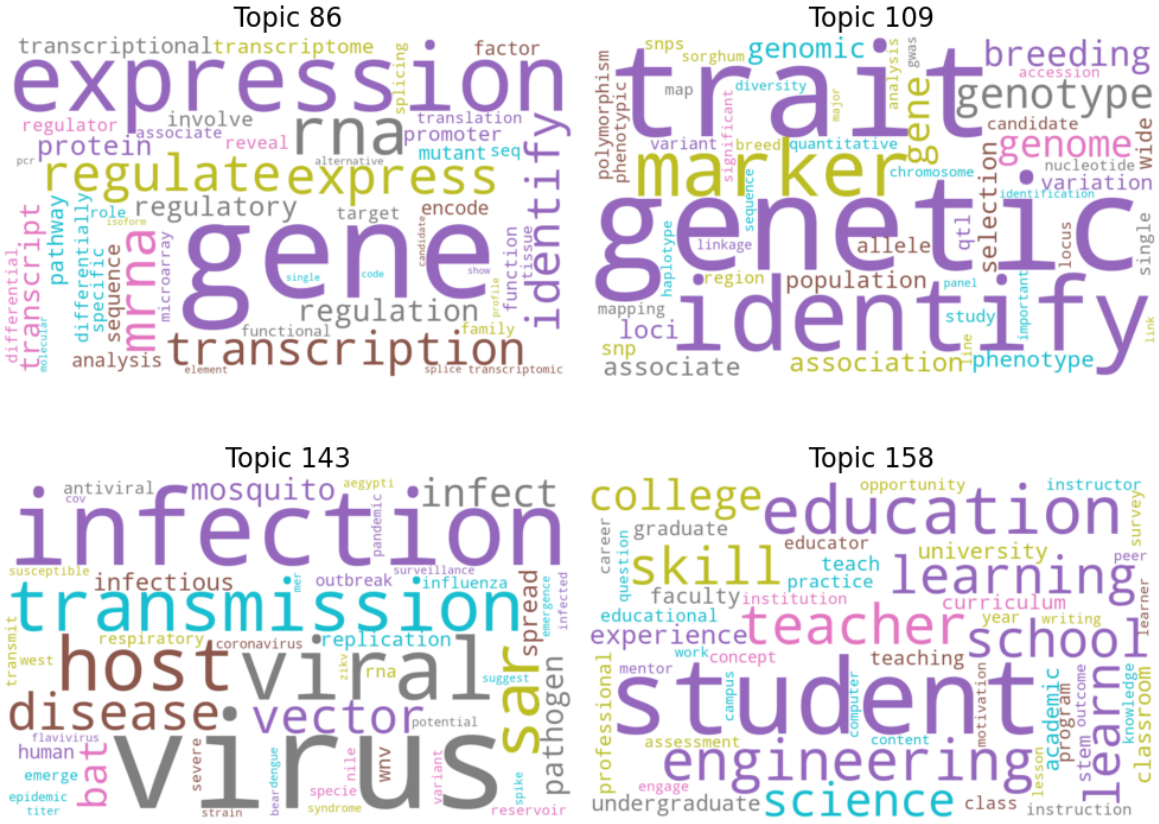


Figure 4.4: Example topics from the LDA model on the combined MWC dataset

and host. Topic 158 contains education-related terms such as student, learning, teacher, and engineering. These examples illustrate that the combined LDA model captures distinct and meaningful research themes relevant to different areas represented in the dataset, providing a suitable foundation for constructing topic-based similarity networks.

4.3.2 Network Construction

A social network can be expressed as a graph $\mathcal{G} = (V, E)$, where V denotes the set of vertices and E the set of edges. In this study, each vertex corresponds to a researcher, and edges represent pairwise relationships based on topic similarity. Constructing a network

in this way requires defining a similarity measure between researchers and then applying edge-weakening strategies to avoid an overly dense graph that would obscure community structures.

Defining Edge Relationships To construct the topic-based collaboration network, we first derived topic probability distributions for each researcher. Using the trained LDA model, topic distributions were initially inferred for individual publications and then aggregated at the researcher level by concatenating all publications of a given scholar into a single document. This approach was chosen deliberately: simply averaging or normalizing document-level topic distributions can distort the underlying probability structure, since the resulting vector may no longer represent a valid distribution over topics. In contrast, treating each researcher's combined publications as a single document allows the LDA inference process to produce a consistent, normalized probability distribution that more accurately reflects the researcher's overall thematic focus. Querying the model on these aggregated texts thus yielded topic distributions that capture the relative research emphasis of each scholar across all discovered topics.

These researcher-level topic distributions serve as high-dimensional profiles that capture the thematic composition of each scholar's work. For instance, a researcher specializing in machine learning may assign higher probability mass to topics related to algorithms, prediction, and data analysis, whereas an ecology researcher will show stronger emphasis on topics such as biodiversity and environmental modeling. Importantly, researchers from different disciplinary backgrounds can still exhibit similar topical signatures if their underlying research themes overlap, forming the basis for potential cross-domain collaboration.

To translate these topic distributions into network edges, we measured pairwise similarity between researchers. Although cosine similarity is one of the most widely used measures for comparing vector representations in information retrieval and text mining

applications [43, 108, 112], it is not ideally suited for probability distributions such as those produced by LDA. Cosine similarity treats vectors geometrically and does not account for the probabilistic nature of topic weights, potentially overstating similarity when two distributions share sparse but non-overlapping mass.

Instead, metrics designed to compare probability distributions provide a more meaningful notion of similarity in this context. The Kullback–Leibler (KL) divergence [58] is a well-established measure for quantifying how one probability distribution diverges from another. It is defined as

$$D_{KL}(\mathbf{P}||\mathbf{Q}) = \sum_i P_i \log\left(\frac{P_i}{Q_i}\right).$$

However, KL divergence is asymmetric and unbounded, making it unsuitable for undirected network construction where pairwise similarity must be reciprocal. To address these limitations, we employed the Jensen–Shannon (JS) divergence [70], a symmetric and bounded variant of KL divergence given by

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

where $\mathbf{M} = \frac{1}{2}(\mathbf{P} + \mathbf{Q})$. Lower JS values indicate greater similarity between two topic distributions. Finally, we defined the edge weight between researchers as $1 - D_{JS}$, such that values closer to one represent stronger topical similarity. This formulation yields an undirected, weighted network that reflects thematic alignment among researchers within and across institutions.

Edge Weakening A direct consequence of using divergence-based similarity is that every researcher pair receives a nonzero similarity score, producing a fully connected network for each dataset. Across the four networks (MSU, WSU, CSU, and the combined MWC), this results in extremely dense structures where each researcher connects to every other

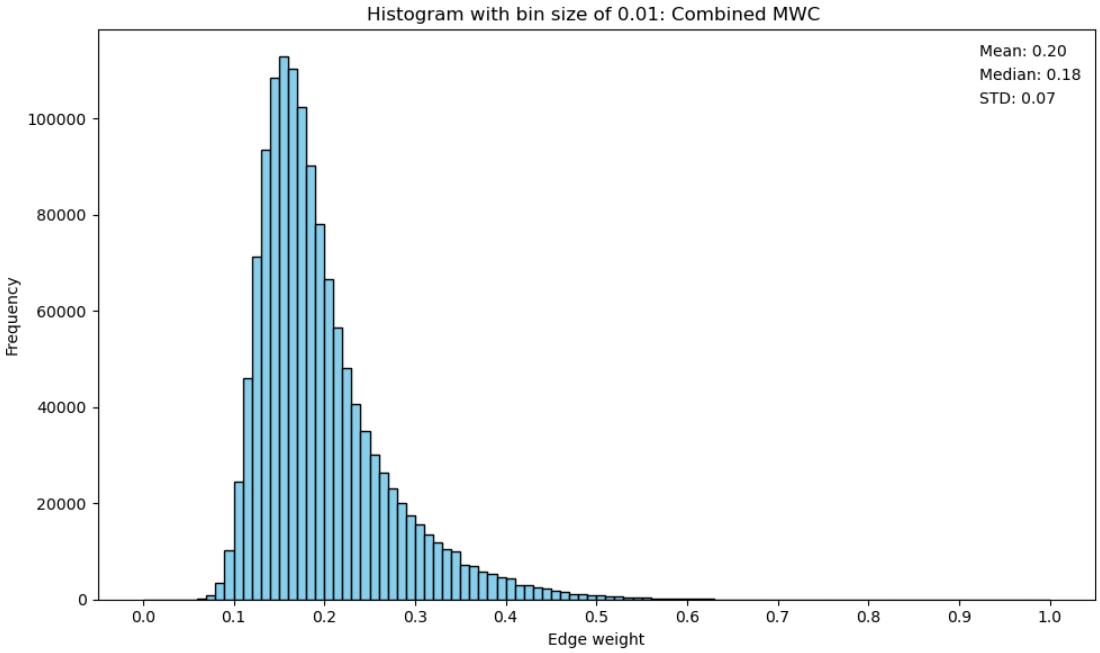


Figure 4.5: Distribution of edge weights in the combined MWC

researcher. For example, in the combined MWC dataset with 1,634 researchers, this yields $\binom{1634}{2}$ = 1,334,161 edges. Such density conceals the modular structure of the network, as every vertex appears uniformly connected, making it difficult for community detection algorithms such as Louvain or Spectral clustering to identify distinct subgroups. These algorithms perform best on sparse graphs, where intra-community connections are relatively strong compared to inter-community ones.

Figure 4.5 shows the empirical distribution of edge weights derived from the Jensen–Shannon similarity scores for the combined MWC network, which reflects the general trend observed across all datasets. The distribution is heavily right-skewed, with most edges concentrated between 0.10 and 0.20, and a mean similarity of 0.20. This indicates that the majority of researcher pairs share only weak topical similarity, while only a small subset of pairs exhibit strong thematic overlap.

To enhance interpretability and improve the effectiveness of community detection,

we applied an "edge-weakening" process that incrementally prunes low-similarity edges. Thresholding is a simple yet effective strategy for highlighting strong topical associations while suppressing weak, noisy links. However, high thresholds can fragment the graph into disconnected components, which is problematic for algorithms such as Spectral clustering that assume a connected Laplacian matrix for valid eigen-decomposition [121]. Louvain clustering, by contrast, can operate on disconnected components, but such fragmentation complicates direct comparisons across thresholds and datasets.

To preserve network connectivity while still enforcing sparsity, we integrated a minimum spanning tree (MST) structure [57, 94] into the thresholding process. The MST was first extracted from the fully connected graph to ensure that all vertices remain reachable through at least one minimal-weight path. As the similarity threshold increased, any vertices or subgraphs that became disconnected were reconnected using the single MST edge of smallest weight bridging the isolated component. This hybrid approach maintains global connectivity required for spectral-based methods while preserving the interpretability benefits of threshold-based sparsification.

We generated a total of eighty network variants by incrementally increasing the similarity threshold from 0.00 to 0.80 in steps of 0.01, applying the MST-based reconnection procedure at each step to preserve global connectivity. For interpretability, Table 4.1 reports representative thresholds at 0.10 intervals. As the threshold rises, weaker links are pruned, yielding progressively sparser but still connected networks that highlight only the most topically aligned researcher pairs. The MST integration ensures that all vertices remain reachable even at higher thresholds, maintaining the spectral clustering requirement of a single connected component while preserving Louvain's ability to reveal modular structure.

Table 4.1: Number of edges retained at different threshold values for the combined MWC

Threshold	0.00	0.10	0.20	0.30	0.40	0.50
#Edges	1334161	1319579	481835	115802	25990	4339

4.3.3 Community Detection

After constructing the scholarly social network with topic-based similarity between researchers, the next step was to detect communities sharing similar topics of interest. The goal is to use these discovered communities to suggest potential collaboration recommendations. Members within a community share topical interests, and the edge weights can be used to rank potential collaborators. Moreover, the discovered communities may provide insights into potential interdisciplinary research centers composed of members from different fields.

To perform community detection, we employed two well-established algorithms suitable for weighted graphs: the Louvain algorithm and Spectral clustering. Both methods produce discrete community assignments, where each vertex belongs to a single community. We used the modularity metric to assess the quality of the discovered communities, measuring the strength of intra-community connections relative to inter-community ones.

Louvain Algorithm The Louvain algorithm, introduced by Blondel et al. [9], is a greedy optimization-based community detection method that partitions a network to maximize modularity. The algorithm operates in two iterative phases: "local modularity optimization" and "community aggregation". Initially, each vertex is placed in its own community. Then, for every vertex, the algorithm evaluates modularity gains by temporarily moving it into the community of each neighbor. The vertex is permanently reassigned to the neighboring community that yields the largest modularity increase. Once no further modularity gain is possible, all vertices within each community are merged into a "super vertex", forming a

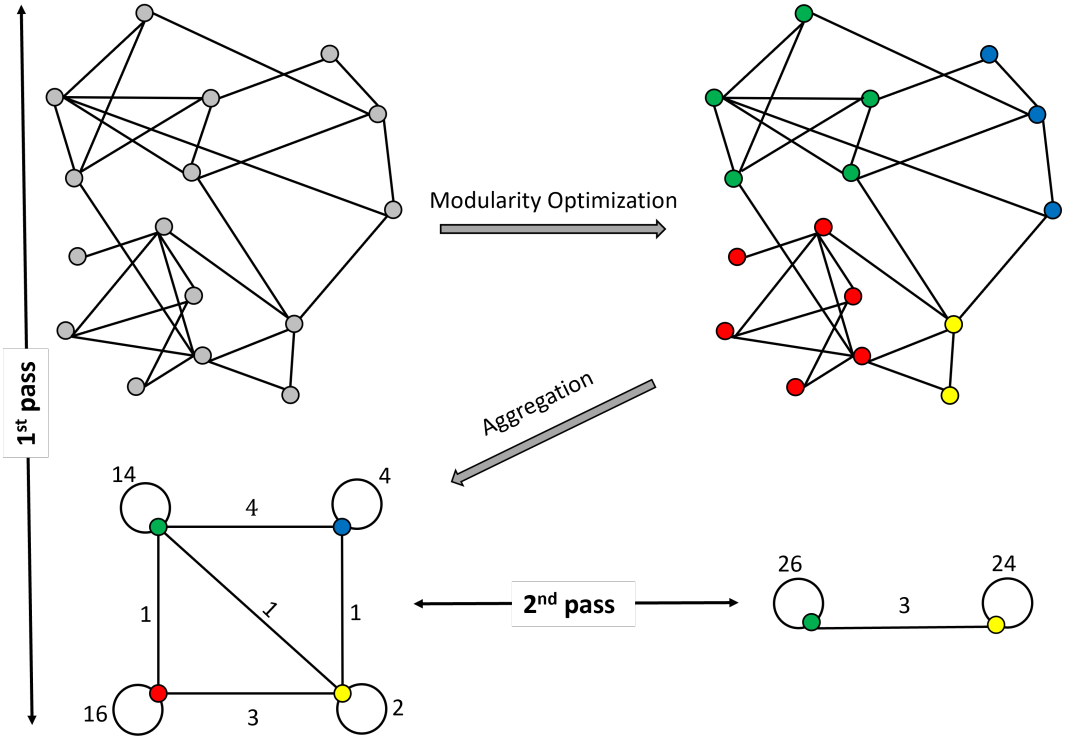


Figure 4.6: Steps followed by the Louvain algorithm

new, smaller network. These two phases repeat until the modularity no longer improves. This hierarchical agglomerative process makes the Louvain algorithm both computationally efficient and suitable for large-scale weighted networks such as ours. Figure 4.6 illustrates the main steps, and Algorithm 4.1 provides the pseudocode.

The Louvain algorithm was particularly appealing for this study because it automatically determines the number of communities that yield the highest modularity, eliminating the need to specify this parameter *a priori*. This property also made it a convenient baseline for comparison with spectral clustering.

Spectral Clustering Spectral clustering [85] is another widely used method for community detection that leverages the eigenstructure of a graph Laplacian. Given a weighted adjacency matrix $\mathbf{A}$, the diagonal degree matrix $\mathbf{D}$ is defined such that $\mathbf{D}_{ii}$ equals the degree of vertex i

Algorithm 4.1 Louvain Algorithm

Initialization:
for each vertex i **do**
    Assign i to its own community
while changes in modularity are significant **do**
    for each vertex i **do**
        for each neighbor community c of i **do**
            Remove i from its current community and place it in c
            Calculate change in modularity ΔQ
        Move i to community with maximum ΔQ
    Merge communities into super vertices
Output: Final community structure maximizing modularity

Algorithm 4.2 Spectral Clustering for Community Detection

Require: Graph $\mathcal{G}(V, E)$, number of communities k
Ensure: Community assignments
Compute the symmetric normalized Laplacian $\mathcal{L}_{\text{sym}}$ from $\mathcal{G}$
Compute the k smallest non-zero eigenvectors of $\mathcal{L}_{\text{sym}}$ to form matrix $U \in \mathbb{R}^{n \times k}$
Normalize the rows of U to unit length
Apply k-means to the rows of U
return Community assignments

and $\mathbf{D}_{ij} = 0$ for $i \neq j$. The unnormalized graph Laplacian is then $\mathcal{L} = \mathbf{D} - \mathbf{A}$. A symmetric normalized version is often preferred:

$$\mathcal{L}_{\text{sym}} = \mathbf{D}^{-\frac{1}{2}} \, \mathcal{L} \, \mathbf{D}^{-\frac{1}{2}},$$

which prevents high-degree vertices from dominating the clustering outcome. The eigenvectors corresponding to the k smallest non-zero eigenvalues of $\mathcal{L}_{\text{sym}}$ form a low-dimensional embedding of the graph, and applying a clustering algorithm (such as k-means) to this embedding partitions the vertices into k communities. Algorithm 4.2 summarizes this process.

Unlike the Louvain algorithm, spectral clustering requires the number of communities k to be specified in advance. To maintain comparability, we used the number of communities

discovered by the Louvain algorithm as the k value for spectral clustering. This ensured that both methods operated on a consistent partitioning scale.

Comparison and Evaluation To evaluate the community detection results, we used the modularity metric to assess intra-community cohesiveness and inter-community separation. Additionally, we computed the Jaccard similarity score $J(\mathbf{A}, \mathbf{B})$ between corresponding communities obtained from the two algorithms:

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

Here, $\mathbf{A}$ and $\mathbf{B}$ represent the vertex sets of two communities. A score of one indicates identical membership, while zero denotes no overlap. We calculated pairwise community similarities between Louvain and spectral results, aligned the most similar pairs, and averaged their scores to obtain an overall community alignment measure. Identical partitions would yield an average Jaccard similarity of one.

4.4 Result and Discussion

To evaluate the effects of the edge threshold across all four institutional networks, we varied the threshold as described in Section 4.3.2 and applied the same community detection pipeline to MSU, WSU, CSU, and MWC. For each dataset, Louvain clustering was run first to obtain the number of communities and the corresponding modularity values, and this community count was then used as the prespecified k for Spectral clustering so that the two methods could be compared on the same number of partitions. Figure 4.7 summarizes the modularity trends for all four institutions in a common 0.30–0.70 range of edge threshold.

Across all institutions, thresholds below 0.30 produced very dense graphs, which resulted in only a few large communities and correspondingly low modularity values. This

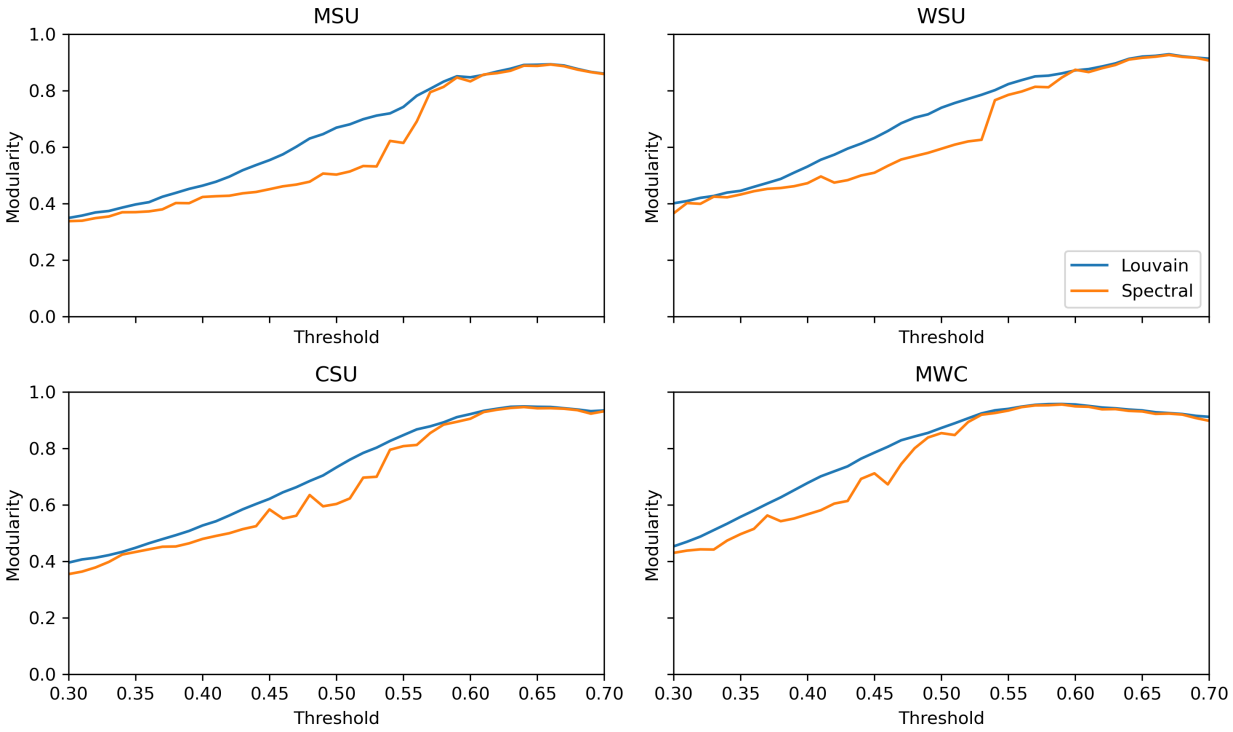


Figure 4.7: Modularity versus edge threshold for the four institutional networks

flat region in the curves indicates that removing only the weakest edges does not yet expose meaningful community structure. Once the threshold entered the 0.30–0.60 range, modularity rose steadily for every dataset, which shows that pruning lower similarity edges helped separate topic-coherent groups. MSU, WSU, and CSU all reached their highest or near-highest modularity values toward the upper part of this interval (around 0.60–0.65), while MWC, being the largest and most topically diverse network, stabilized slightly earlier, near 0.55–0.58. Beyond 0.70, modularity changes became small and in some cases declined, indicating that further pruning would begin to fragment the networks. In all cases, Louvain modularity remained equal to or higher than Spectral modularity, as expected given that Louvain directly optimizes the modularity objective. As the edge threshold increased and the networks became sparser, the difference between the two methods narrowed, with Spectral producing modularity values closer to those of Louvain in the sparser graphs.

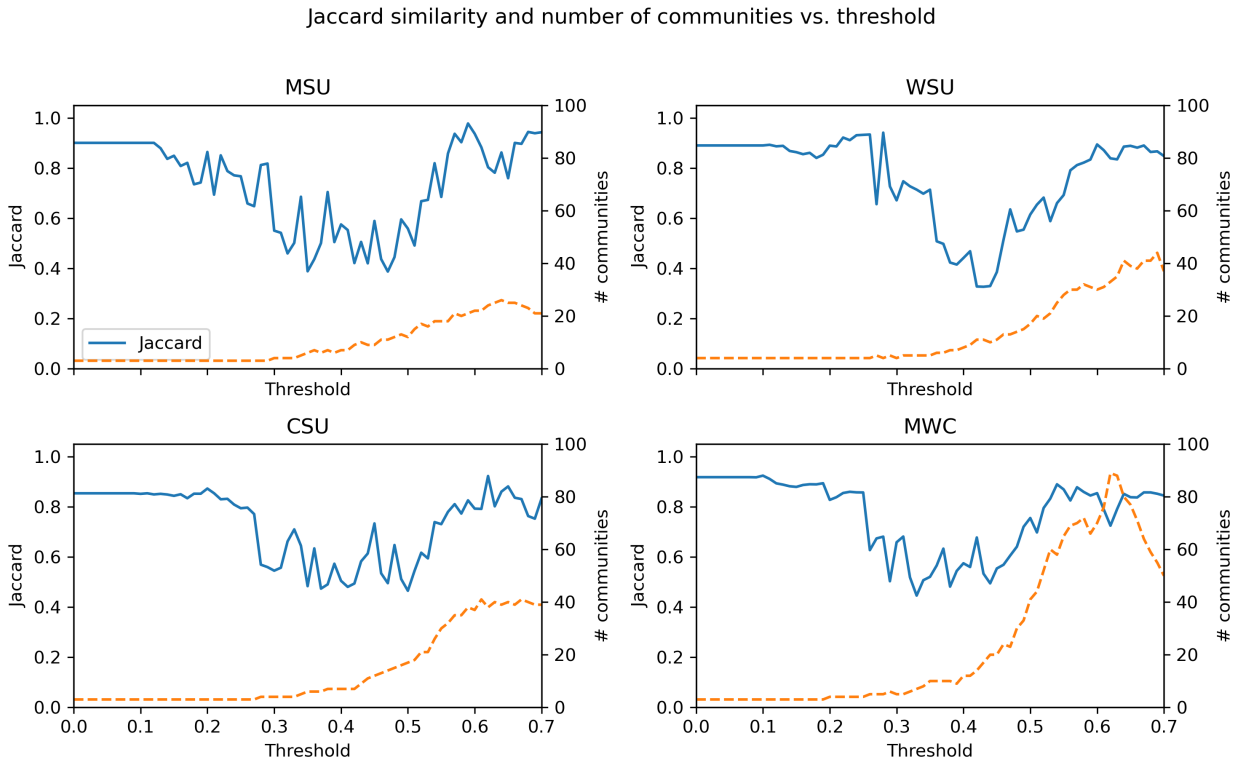


Figure 4.8: Louvain–Spectral Jaccard similarity and number of communities across edge thresholds for the four institutional networks

Following the modularity comparison, we examined the agreement between Louvain and Spectral clustering by measuring the average Jaccard similarity of their detected communities across thresholds. Figure 4.8 presents the Jaccard similarity (left Y-axis) and the corresponding number of communities (right Y-axis) for all four institutions. At very low thresholds, the networks remain highly connected, resulting in only a few large communities and near-perfect similarity (Jaccard ≈ 1.0) between both algorithms. As the threshold increases and weak edges are pruned, the number of detected communities rises sharply, while Jaccard similarity temporarily drops, reflecting small partitioning differences between the modularity-based and eigenstructure-based methods. Beyond the mid-range (around 0.5–0.6), the Jaccard curves rise again as both algorithms converge on comparable community boundaries in the sparser, more modular network structures.

To determine the most representative network configuration for each institution, we jointly considered the modularity trends and the Jaccard similarity patterns. A suitable threshold should preserve sufficient connectivity to maintain interpretable community structures while maximizing modularity and agreement between Louvain and Spectral clustering. Extremely low thresholds yielded overly dense networks with few broad communities, whereas very high thresholds fragmented the graphs and reduced interpretability. Therefore, we selected the threshold region that balanced both criteria—where modularity values were near their peak and Jaccard similarity began to rise again, indicating convergence between the two methods. For most institutions, this range occurred between 0.6 and 0.65, reflecting a stable and well-separated network structure. Accordingly, we next focus on the MSU network as a representative case, using its optimal threshold to present detailed community statistics and qualitative insights. Supplementary results for WSU, CSU, and the combined MWC network are provided in Appendix A.

Table 4.2 summarizes how the MSU network structure changes as we increase the edge threshold. Columns show the number of detected communities (Comm), average Jaccard similarity (AvgJ), and the community size distribution: maximum (Max), minimum (Min), median (Med), and standard deviation (Std) for Louvain (L) and Spectral (S) partitions. At low thresholds (0.00 and 0.12) the network is almost fully connected, resulting in only three very large communities and high agreement between Louvain and Spectral (AvgJ = 0.90), but these partitions are too coarse for recommendation. As the threshold increases to 0.24–0.48, the number of communities grows and the community sizes become more balanced, although the Jaccard score temporarily drops because the two algorithms start splitting the network differently. The configuration at **0.57** provides the best overall balance: both algorithms agree almost perfectly (AvgJ = 0.94), the number of communities (21) is suitable for qualitative inspection, and the median community size (11 members) is large enough to be meaningful but not so large that topics are blended. Modularity at this level also

Table 4.2: Louvain and Spectral community statistics on MSU across selected thresholds

TH	**Comm**	**AvgJ**	**Max(L)**	**Min(L)**	**Med(L)**	**Std(L)**
			Max(S)	**Min(S)**	**Med(S)**	**Std(S)**
0.00	3	0.90	111	55	110	26.17
			111	63	102	20.83
0.12	3	0.90	111	55	110	26.17
			111	63	102	20.83
0.24	3	0.77	108	61	107	21.92
			126	60	90	26.98
0.36	7	0.44	91	3	51	30.99
			106	21	26	28.68
0.48	12	0.44	54	3	25.5	16.38
			110	7	14.5	27.08
0.57	**21**	**0.94**	36	3	11	9.07
			36	3	11	9.86
0.63	25	0.78	50	2	5	11.98
			50	3	7	11.12
0.70	21	0.94	76	2	5	17.45
			76	2	5	17.52

remains high (0.81 for Louvain and 0.79 for Spectral), only slightly lower than the maximum observed at 0.63 (0.88), but without the over-fragmentation seen at higher thresholds. Higher thresholds such as 0.63 and 0.70 either fragment the network into smaller groups (median = 5) or produce highly uneven community sizes, even if the Jaccard score remains high. Therefore, we used the 0.57 configuration for the MSU community-level analysis presented

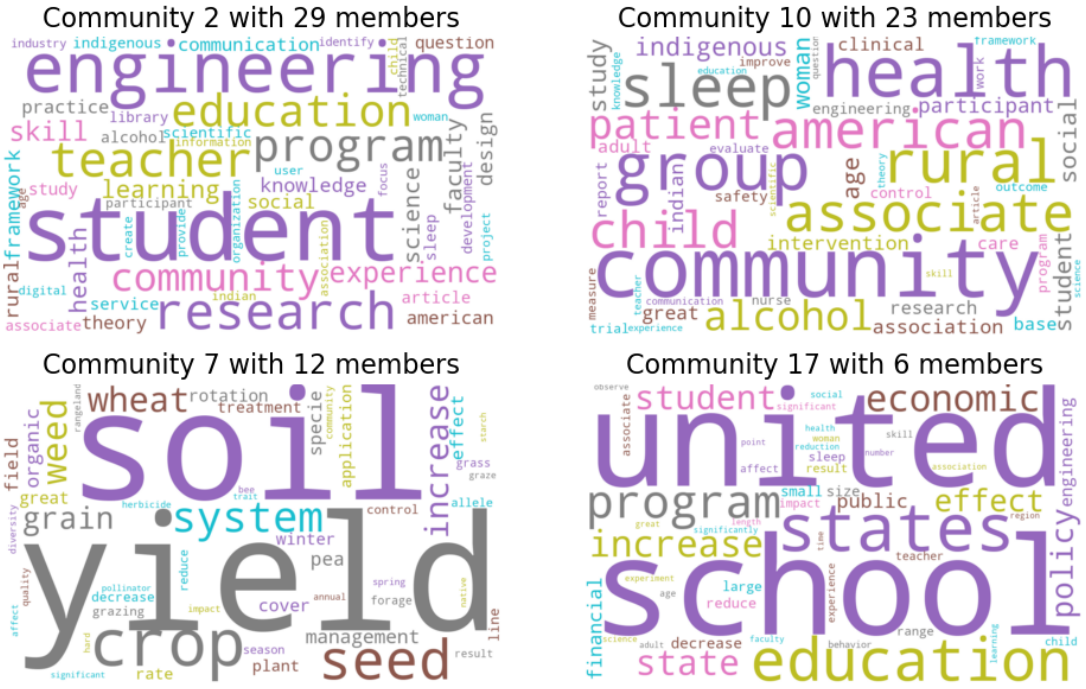


Figure 4.9: Representative MSU Louvain communities at the selected threshold

next. The same selection procedure was applied to the WSU, CSU, and MWC networks, and their corresponding threshold statistics are included in Appendix A.

After identifying the optimal threshold, we examined the community-level structures and their topical coherence. Each community was analyzed by aggregating all publications from its members and querying the trained LDA model to extract the dominant themes. Figure 4.9 illustrates four representative MSU communities at the 0.57 configuration, selected to showcase the diversity of network structures and thematic focus. These four examples were chosen simply to show a mix of community sizes and topic areas, starting with a larger and more mixed group and then moving toward smaller communities that have a tighter departmental focus.

Community 2 is the largest of the examples, comprising 29 members across thirteen departments, including Education, Engineering, Political Science, and Health & Human

Development. Its wordcloud reveals terms such as “student”, “teacher”, “program”, and “engineering”, suggesting a multidisciplinary blend of STEM and social science researchers focused on education, outreach, and research training. Community 10 includes 23 members primarily from Health & Human Development, Psychology, and Nursing. Prominent words like “sleep”, “patient”, “rural”, and “community” indicate a cohesive health-focused cluster with strong interdisciplinary connections between clinical and behavioral research areas.

Community 7, with 12 members, centers on agricultural and environmental topics, dominated by Land Resources & Environmental Sciences and Plant Sciences. Its key terms—“soil”, “yield”, “weed”, and “rotation”—reflect applied agricultural research tied to sustainable land management. Finally, Community 17, a compact group of six researchers from Agricultural Economics, Business, and Sociology & Anthropology, exhibits a distinct focus on economic and educational policy, as seen through terms like “united”, “states”, “school”, and “program”. Together, these examples demonstrate that the 0.57 threshold not only preserves meaningful community structure but also captures topical coherence aligned with departmental expertise, confirming that the selected configuration yields interpretable and functionally relevant research clusters within the MSU network.

The structural statistics in Table 4.3(a) provide a detailed view of how each community is organized and how collaboration currently moves within it. The column labeled “Internal” counts all topic-based similarity edges among members of a community, representing the full set of topical relationships available within that group. We then use OpenAlex coauthorship data to identify which of those internal edges correspond to actual prior collaborations; these appear under the “Coauthored” column. Subtracting the coauthored edges from the total internal edges yields the value reported as “Potential,” which reflects pairs of researchers who are topically related but have not previously published together. This distinction is important because it separates existing collaboration patterns from the additional opportunities that the topic-based structure reveals.

Table 4.3: Community-level structural and departmental statistics for MSU

Community	**Size**	**Internal**	**Coauthored**	**Potential**
2	29	112	16	96
10	23	50	10	40
7	12	21	19	2
17	6	11	3	8

(a) Structural statistics for selected MSU communities

Community	**Dominant Departments (count)**
2	Education (8), Mathematical Sciences (4), Political Science (3), Health & Human Development (3), others
10	Health & Human Development (7), Psychology (5), Sociology & Anthropology (4), Nursing (4), others
7	Land Resources & Environmental Sciences (9), Plant Sciences (2), Animal & Range Sciences (1)
17	Agricultural Economics (4), Business (1), Sociology & Anthropology (1)

(b) Departmental composition of selected MSU communities

Across the four example communities, many internal connections do not correspond to prior coauthorship, which suggests that several collaboration opportunities remain unrealized even among researchers who work on closely related themes. The departmental composition in Table 4.3(b) reinforces this point: many of the potential collaborations link researchers from different academic units rather than within a single department. For instance, Community 2 includes members from Education, Engineering, Political Science, and Health & Human Development, meaning that most of its potential collaborations cross departmental

boundaries. This pattern directly supports our hypothesis that topic-defined communities capture cross-domain affinities that traditional coauthorship networks overlook. Even before applying any ranking or recommendation algorithm, the community structure itself reveals meaningful interdisciplinary opportunities that are embedded within the MSU research landscape but not yet reflected in its publication record.

Overall, the MSU results show that a topic-based network, once thresholded appropriately, produces communities that reflect both the thematic structure of the institution and the collaboration gaps that exist within it. The examples examined here illustrate how researchers with shared topical interests often remain unconnected in the coauthorship record, even when they span multiple departments or participate in related research areas. These patterns suggest that topic-defined communities capture the underlying structure of scholarly activity more fully than collaboration histories alone, providing a clearer picture of where new connections could naturally emerge. Although MSU is used as the detailed case study in this section, similar behaviors were observed in the WSU, CSU, and MWC networks, indicating that this community detection framework reveals comparable structural and topical relationships across institutions. These observations set the stage for the more advanced analyses in the following chapters, where hierarchical structure, author representation, and model stability are explored in greater depth.

4.5 Summary

This chapter presented a content-based approach to community detection in scholarly networks for identifying potential interdisciplinary collaborations. We constructed a network using topic-based similarity derived from publication metadata, hypothesizing that shared topical interests could bridge disciplinary divides. The resulting communities validated this hypothesis, revealing clusters containing researchers from diverse academic backgrounds.

Applying a threshold-based edge weakening strategy produced a more meaningful

network structure with improved modularity and interpretability. Both Louvain and Spectral clustering performed comparably, with Louvain consistently achieving slightly higher modularity. Qualitative inspection confirmed that the discovered communities reflected logical topical and departmental groupings, with several communities revealing promising cross-disciplinary connections.

While this chapter focused on discrete community detection within a flattened network structure, real-world scholars often contribute to multiple, nested research areas. Building on these insights, subsequent chapters extend this framework to explore hierarchical community organization, address publication imbalance through author representation strategies, and evaluate the stability and interpretability of topic-based collaboration recommendations. Together, these extensions advance the overall goal of developing a scalable and explainable system for identifying interdisciplinary research opportunities across institutions.

In parallel, future extensions of this work will consider localized network construction and incremental updates to improve scalability across larger, cross-institutional datasets, preserving the ability to surface new interdisciplinary collaborations as the scholarly landscape evolves.

CHAPTER FIVE

HIERARCHICAL COMMUNITY MODELING

In this chapter, we address the second research question of this dissertation: "How can hierarchical community detection capture multi-level structures in scholarly networks and improve our understanding of topic-based collaboration patterns?" While the previous chapter focused on discovering communities within a single layer of topic-based connections, real-world scholarly networks often exhibit multiple levels of organization. Smaller, specialized research groups may exist within broader disciplinary clusters, reflecting the nested nature of academic domains and research themes. Understanding this hierarchy is important for modeling how collaborations form and evolve at different levels of granularity—from tightly connected research teams to higher-level disciplinary alignments.

The chapter begins with the motivation and problem statement, followed by a review of related works on hierarchical community detection. We then describe the methodology, including the hierarchical extensions of Spectral and Louvain algorithms, and introduce the proposed Nested Hierarchical Louvain (NH-Louvain) model. Next, we present the evaluation framework based on the Cophenetic Correlation Coefficient (CPCC) and outline the experimental design conducted on both the topic-based scholarly network and several synthetic networks. Finally, we discuss the results and conclude with observations and directions for future work.

5.1 Motivation and Problem Statement

Community detection plays a central role in social network analysis by identifying groups of vertices that are more densely connected to each other than to the rest of the network. However, most conventional algorithms produce a single-layer partition of the

network, assuming that each vertex belongs to one community. This assumption often fails to capture the complex, multi-level organization seen in real-world systems. For instance, in academic environments, researchers may belong to smaller research groups that share specific topics, while these groups themselves belong to larger disciplinary or institutional clusters. Such relationships naturally form a hierarchy that mirrors both specialization and breadth within the scholarly landscape.

In Chapter 4, we constructed a scholarly social network based on topic similarity derived from publication metadata to uncover potential interdisciplinary collaborations. That approach emphasized distinct, non-overlapping communities using algorithms such as Louvain and Spectral clustering. Yet, the topical structure underlying the network often follows a hierarchy: "machine learning" and "computer vision" may both belong under the broader theme of "artificial intelligence", which in turn aligns with the larger discipline of "computer science". If topics are hierarchically organized, the researcher network constructed from those topics is also likely to exhibit hierarchical community patterns.

Recognizing these multi-level relationships can provide a more comprehensive understanding of collaboration structures and improve the foundation for recommendation models that operate across different topical scales. For example, identifying hierarchical communities allows us to recommend not only direct peers within a subfield but also potential collaborators who share higher-level research goals. Thus, our focus in this chapter is to model and evaluate the hierarchical organization embedded in the topic-based network.

To explore this idea, we propose two approaches for hierarchical community detection. First, we extend the traditional Louvain algorithm to a "Nested Hierarchical Louvain" (NH-Louvain) framework that recursively partitions subgraphs to uncover deeper levels of community structure. Modularity-based methods are known to favor larger partitions and can therefore overlook small but meaningful groups, a limitation often referred to as the resolution limit. This recursive process helps reduce the effect of this limitation by

checking community structure again inside each group, which makes it easier for smaller subcommunities to appear when they are present. Second, we develop a hierarchical variant of Spectral clustering by integrating Hierarchical Agglomerative Clustering (HAC) on the spectral embeddings of the network. While Spectral clustering is generally applied for flat community detection, this adaptation enables it to identify communities at multiple levels of granularity. To the best of our knowledge, prior studies have rarely combined HAC with Spectral clustering for hierarchical community detection within social networks, making this an additional methodological contribution of this work.

Furthermore, to evaluate the consistency and quality of the detected hierarchies, we employ the Cophenetic Correlation Coefficient (CPCC), a metric traditionally used in hierarchical clustering. We adapt CPCC for hierarchical community analysis to measure how well each algorithm preserves the original pairwise relationships between researchers across different hierarchical levels. Although CPCC is not commonly used in community detection, we show that it can serve as a useful complementary measure to modularity, providing quantitative insight into the depth and coherence of the resulting hierarchies.

Problem Statement: Given a topic-based scholarly network derived from researchers' publication metadata, how can hierarchical community detection methods uncover multi-level structures that reveal both fine-grained and higher-level collaborations?
Hypothesis: We hypothesize that applying a nested hierarchical community detection framework, through both NH-Louvain and Spectral–HAC, will identify deeper and more interpretable community structures compared to traditional flat approaches. Such hierarchical modeling is expected to yield a more nuanced view of the scholarly landscape, highlighting connections across multiple topical layers and providing a stronger analytical foundation for future recommender system design.

5.2 Related Work

Community detection has long been a central problem in social network analysis, aiming to identify cohesive groups of vertices within complex systems. Traditional algorithms such as the Louvain method [9] and Hierarchical InfoMap [105] have been used widely to detect modular structures in large networks. However, these methods typically produce only a single-layer partition and therefore miss finer subdivisions that may exist within broader communities. Louvain uses a multilevel optimization process, but the intermediate levels created during this procedure are not meaningful community layers and are discarded once modularity is maximized. As a result, the method cannot represent nested community structure. Moreover, Louvain is affected by the resolution limit, where smaller but meaningful groups are merged into larger ones because their separation yields only a small modularity gain [29]. This is problematic in topic-based scholarly networks, where such small groups often correspond to emerging or specialized research themes. Similarly, InfoMap recursively partitions a network based on information flow, but it can be sensitive to parameter choices and often performs poorly on dense, weighted networks like our topical similarity–based scholarly network, where flow patterns are not clearly separated.

Several extensions have been proposed to better capture hierarchical organization in networks. Clauset et al. [18] developed one of the earliest modularity-based hierarchical approaches for large networks. Li et al. [65] introduced a recursive partitioning framework that formalizes hierarchical community detection with statistical guarantees, offering deeper insight into multi-level network structures. Lancichinetti et al. [61] proposed a hierarchical benchmark model that remains widely used for evaluating hierarchical community detection algorithms. Peixoto et al. [92] presented the hierarchical stochastic block model (hSBM), a probabilistic formulation that infers community hierarchies by fitting generative models to network data. Although these model-based methods provide statistical rigor, their

computational complexity increases rapidly with network size, making them less suitable for large-scale scholarly networks constructed from content-based similarity.

In the spectral domain, Wahl et al. [122] introduced a hierarchical fuzzy spectral clustering framework that applies fuzzy c-means to spectral embeddings across multiple values of k, linking the resulting partitions through Jaccard similarity to construct a hierarchical structure. This approach produces soft, overlapping communities, which may be advantageous in settings where researchers contribute to multiple topical areas. However, running fuzzy c-means repeatedly over multiple cluster counts introduces substantial computational cost, making the method less practical for large, dense networks such as topical similarity–based scholarly graphs. Our Spectral–HAC approach avoids this overhead by building a hard hierarchy through a single agglomerative process. Although we do not address overlapping community detection here, fuzzy spectral methods highlight a promising direction for future work.

Beyond modularity and model-based formulations, deep learning has recently been adopted for hierarchical community detection. Ding et al. [24] introduced a self-evolving hierarchical community detection framework using deep neural networks to learn latent hierarchical representations. While these approaches can capture complex structures automatically, they often require large training datasets, careful parameter tuning, and labeled supervision, which are rarely available in scholarly settings. Moreover, the resulting hierarchies are encoded within embedding spaces, making them less interpretable compared to explicit, structure-based partitions.

In contrast, the approach presented in this dissertation focuses on transparent, unsupervised algorithms that reveal interpretable hierarchies directly from network structure. The proposed NH-Louvain algorithm recursively partitions detected communities to overcome modularity's resolution limit and expose finer layers of organization. Additionally, we extend Spectral clustering into a hierarchical variant by applying HAC on spectral embeddings,

enabling multi-level detection through eigenstructure analysis—a combination that, to our knowledge, has not been previously applied in social network community detection. Finally, we adopt the CPCC as a quantitative evaluation measure for hierarchical structures, providing a novel metric for assessing how well community hierarchies preserve the original pairwise relationships among vertices. Although CPCC is traditionally used in hierarchical clustering, its adaptation here offers a meaningful and interpretable way to evaluate the coherence and depth of network hierarchies.

5.3 Methodology

In this section, we describe the methodological framework used to identify hierarchical community structures in the topic-based scholarly network. Our goal is to investigate how different hierarchical community detection strategies can reveal the multi-level organization of research communities that emerge from topic similarities among scholars. We employ three algorithms: the standard Louvain algorithm, which serves as the baseline; the NH-Louvain; and, finally, the Spectral–HAC. We then evaluate the hierarchical quality of each algorithm using the CPCC, a structural measure that quantifies how closely the resulting dendrogram reflects the pairwise similarities in the original network. The final part of this section outlines the experimental design for both the scholarly and synthetic datasets used in our analysis.

5.3.1 Louvain (Baseline)

The Louvain algorithm [9] serves as the foundation for our hierarchical analysis. As discussed in Section 4.3.3, Louvain is a greedy modularity optimization algorithm that iteratively merges vertices or communities to maximize modularity, a measure of the density of edges within communities relative to a null model of random connections. This optimization proceeds in two main phases: a local modularity optimization phase, where each

vertex is reassigned to the community yielding the highest modularity gain, and a network aggregation phase, where each community is treated as a single super-vertex, producing a successively coarser representation of the network. Through this iterative process, Louvain implicitly generates a hierarchy of communities, although in practice this hierarchy is often shallow due to the resolution limit problem discussed later. The algorithm's scalability and unsupervised nature make it an ideal starting point for building a more expressive hierarchical framework. We next describe how we extend this baseline into a divisive, multi-level procedure.

5.3.2 Nested Hierarchical Louvain

While Louvain is capable of revealing hierarchical community structures, its depth of hierarchy is often limited. This limitation stems from its reliance on modularity maximization, which tends to favor larger communities. To overcome this bias and to reveal finer substructures within the network, we introduce the NH-Louvain algorithm. NH-Louvain follows a top-down strategy inspired by traditional divisive hierarchical clustering [46] and can be summarized as follows:

- **Initialization:** Begin with all data points assigned to a single cluster.
- **Divisive Step:** The selected cluster is divided into two or more subclusters based on a specific criterion.
- **Recursion:** The divisive process is applied recursively to each of the resulting subclusters until a stopping criterion is met.
- **Stopping Criteria:** The algorithm terminates when a stopping criterion is satisfied. This could be a predefined number of clusters or when further splitting is not feasible.
- **Hierarchy Formation:** Throughout the process, a hierarchical structure, often represented as a dendrogram, is constructed.

Algorithm 5.3 Nested Hierarchical Louvain

Require: Graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$

```
1: function NH-Louvain(𝒢)
2:     𝒫 ← Louvain(𝒢)
3:     if |𝒫| ≤ 1 then
4:         return leaf vertex labeled with 𝒫
5:     else
6:         create an internal vertex for 𝒫
7:         for each community C ∈ 𝒫 do
8:             𝒢_C ← 𝒢[vertices in C]
9:             attach NH-Louvain(𝒢_C) as a child
10:        return internal vertex
```

The process starts by placing all vertices in the same community. For the divisive step, we use the base Louvain algorithm to determine the first level of network partitions. We then recursively apply the Louvain algorithm on the partitions obtained in the previous step until reaching a stopping criterion. The stopping condition may occur when the partitions lead to a single vertex or a single community that possibly forms a clique. The process ultimately creates a dendrogram revealing the hierarchical structures present in the network.

Algorithm 5.3 shows the pseudocode of the NH-Louvain algorithm. The algorithm takes the entire network $\mathcal{G}$ as input and runs the base Louvain to obtain the initial partitions at line 2. Then, at line 3, it checks the stopping criterion for the recursive call, verifying whether the partitions obtained from Louvain form a single partition or are empty. If this condition is satisfied, the algorithm assigns it as a leaf vertex of the parent community and terminates at line 4. Otherwise, for each community in the partition, it treats that community as a child of the parent community and recursively calls the function with the subgraph induced by the vertices in that community, effectively shrinking the size of the original network (lines 6–8). Finally, the algorithm returns a dendrogram that encodes the hierarchical relationships among the partitions, where the leaves represent a single vertices or possibly a clique and internal vertices represent higher-level aggregations.

The main challenge with modularity-based approaches is the "resolution limit problem", which causes smaller communities to be merged into larger ones even when they are internally well-defined [29]. This issue arises because modularity compares observed intra-community edges to the expected number of such edges in a random network, scaling the comparison by the total number of edges m in the graph. For an adjacency matrix $\mathbf{A} = [a_{ij}]$ with degrees d_i and community labels c_i, modularity is defined as

$$Q = \frac{1}{2m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(a_{ij} - \frac{d_i d_j}{2m} \right) \delta(c_i, c_j),$$

where $\delta(c_i, c_j)$ equals 1 if vertices i and j belong to the same community and 0 otherwise. Maximizing Q favors partitions that yield large modularity improvements, which often correspond to merging smaller subgraphs whose contribution to the overall modularity is negligible when m is large. Fortunato et al. [29] showed that Louvain fails to detect communities smaller than approximately $\sqrt{4m}$ edges.

Several extensions of modularity introduce a resolution parameter γ to adjust the scale at which communities are detected [97]. The modified objective takes the form

$$Q(\gamma) = \frac{1}{2m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(a_{ij} - \gamma \frac{d_i d_j}{2m} \right) \delta(c_i, c_j),$$

where higher values of γ encourage smaller communities and lower values favor larger ones, partially addressing the resolution limit. However, selecting an appropriate value of γ requires domain knowledge because different settings can lead to substantially different partitions. In addition, the resulting modularity scores are not directly comparable across resolutions, which complicates the evaluation of partition quality. Instead of tuning γ, we adopt a hierarchical approach that examines community structure at multiple levels without relying on scale parameters. This avoids the need to commit to a single resolution and aligns

better with the layered topical organization present in scholarly networks.

In NH-Louvain, the recursive structure helps mitigate this bias. Each recursive call operates on a smaller induced subgraph, effectively reducing the total number of edges m considered during modularity optimization. This reduction tightens the resolution scale, allowing Louvain to identify smaller, more meaningful communities that would otherwise be absorbed into larger ones in a single global pass. Consequently, the resulting dendrogram captures both coarse and fine-grained community structures, improving interpretability and providing a more balanced view of the hierarchical organization of the scholarly network.

5.3.3 Spectral Clustering with Hierarchical Agglomerative Clustering

Spectral clustering is another widely used approach for community detection, particularly suitable for identifying non-convex clusters through the spectral properties of the graph Laplacian. The core idea is to represent the graph as a Laplacian matrix and project the vertices into a lower-dimensional space using the eigenvectors associated with the smallest eigenvalues. As described in Section 4.3.3, the normalized Laplacian is computed as

$$\mathcal{L}_{\text{sym}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{A})\mathbf{D}^{-\frac{1}{2}},$$

where $\mathbf{A}$ is the adjacency matrix and $\mathbf{D}$ is the diagonal degree matrix. After computing the k smallest eigenvectors, we form the spectral embedding matrix $\mathbf{U} \in \mathbb{R}^{n \times k}$, where each row corresponds to a vertex representation in the reduced space. These embeddings capture the intrinsic structure of the network, and clustering them can reveal communities.

Unlike traditional spectral clustering that uses k-means to produce a flat partition, our approach employs HAC [125] to uncover multi-level community structures. HAC successively merges clusters based on a chosen linkage criterion, resulting in a dendrogram that captures relationships between communities at different levels of granularity.

HAC begins with each vertex as its own cluster and iteratively merges the two clusters

with the smallest inter-cluster distance until all vertices belong to a single cluster or a stopping condition is reached. The distance between clusters can be computed in several ways [116]:

- **Single linkage:** uses the minimum pairwise distance between elements of two clusters. It can capture irregular shapes but tends to form long, chain-like structures.
- **Complete linkage:** uses the maximum pairwise distance between clusters. It produces compact and well-separated clusters, which generally yield clearer hierarchical structures.
- **Average linkage:** considers the mean pairwise distance between clusters, offering a balance between the two extremes above.
- **Ward linkage:** merges clusters that lead to the smallest increase in within-cluster variance, making it effective when distances represent squared Euclidean metrics.

The choice of linkage criterion influences the structure of the resulting hierarchy, since each definition emphasizes different aspects of cluster separation. In this work, HAC is presented in its general form, and the selection of a specific linkage function is addressed later in the experimental design.

Algorithm 5.4 presents the pseudocode for applying HAC on spectral embeddings. The algorithm requires the spectral embedding matrix as input, which is obtained from the eigenvectors of the normalized graph Laplacian discussed earlier. At line 2, the pairwise similarities between the embedding vectors are computed using a chosen distance metric. Lines 3–8 describe the agglomerative process: the two closest clusters are identified and merged based on the selected linkage criterion, and the distances are updated accordingly at each iteration. Each merge operation is recorded to construct the dendrogram that captures the complete sequence of hierarchical merges.

Algorithm 5.4 Hierarchical Agglomerative Clustering on Spectral Embeddings

Require: Spectral embedding matrix $\mathbf{U} \in \mathbb{R}^{n \times k}$, distance metric $d(\cdot, \cdot)$, linkage ℓ

Initialize clusters $\mathcal{C} \leftarrow \{\{1\}, \{2\}, \ldots, \{n\}\}$
Compute pairwise distances using $d(\cdot, \cdot)$ on rows of $\mathbf{U}$
while $|\mathcal{C}| > 1$ **do**
    $(C_p, C_q) \leftarrow \arg\min_{C_a \neq C_b} \mathrm{Dist}_\ell(C_a, C_b)$
    Merge C_p and C_q into C_{pq}
    Update distances under linkage ℓ
    Record merge level for dendrogram construction
return dendrogram $\mathcal{T}$ with hierarchical merges

By applying HAC to the spectral embeddings, we obtain a hierarchical partition that can be examined alongside the hierarchy produced by NH-Louvain. Spectral–HAC provides a complementary perspective by deriving its structure from the graph's eigenvectors rather than from modularity-based refinement. In the following section, we describe the evaluation framework used to assess the resulting hierarchies.

5.3.4 Cophenetic Correlation Coefficient

The Cophenetic Correlation Coefficient (CPCC) is a popular metric for evaluating hierarchical clustering in an unsupervised setting [113, 118]. It measures the goodness-of-fit of a hierarchical clustering solution to the original data by quantifying the similarity between the pairwise distances among data points and the distances in the dendrogram produced by the hierarchical clustering algorithm. However, it has not received much attention for evaluating hierarchical community structures in the context of community detection in social networks.

To compute CPCC, we first obtain the cophenetic distance matrix from the dendrogram produced by a hierarchical community detection algorithm. Each vertex in the dendrogram represents a community at that level. The cophenetic distance between two vertices is the distance or height of the lowest common ancestor vertex where the two vertices first merge into a community. Formally, let $h(u)$ denote the height associated with an internal vertex

u in the dendrogram, and let $\mathrm{LCA}(v_i, v_j)$ denote the lowest common ancestor of the leaves corresponding to vertices v_i and v_j in the original network. The cophenetic distance between v_i and v_j is then

$$\delta_{ij} \;=\; h\big(\mathrm{LCA}(v_i, v_j)\big).$$

The cophenetic distance matrix $\mathbf{\Delta} = [\delta_{ij}]$ then contains the cophenetic distances of all vertex pairs in the network derived from the dendrogram.

Next, we compute the pairwise distance matrix for the original data points. As we construct networks based on content data, we obtain an undirected weighted graph where the weights represent the strength of the connection between two vertices. Therefore, for the similarity matrix of the original data points, we use the adjacency matrix of the constructed graph. Let $\mathbf{A}$ denote the adjacency matrix with a_{ij} representing the edge weight between vertices v_i and v_j. Let $\mathbf{\Delta}$ denote the cophenetic distance matrix with $\delta_{i,j}$ representing the cophenetic distance between vertices v_i and v_j. The CPCC between these two matrices is computed as the Pearson correlation between their upper-triangular entries:

$$\mathrm{CPCC} \;=\; \frac{\sum_{i<j}(a_{ij} - \bar{a})(\delta_{ij} - \bar{\delta})}{\sqrt{\sum_{i<j}(a_{ij} - \bar{a})^2}\,\sqrt{\sum_{i<j}(\delta_{ij} - \bar{\delta})^2}},$$

where $\bar{a}$ and $\bar{\delta}$ are the corresponding means. Values range from -1 to 1, with 1 indicating a perfect fit between the dendrogram and the original similarities, 0 indicating no correlation, and -1 indicating perfect negative correlation.

5.3.5 Experimental Design

We evaluate the proposed methods on two types of networks: (1) synthetic networks generated using hierarchical stochastic block models (SBMs) [42] to create controlled hierarchical structures, and (2) a topic-based scholarly network constructed from publication metadata across three universities. The details of each network and the corresponding design

decisions are discussed below.

Synthetic Network To further evaluate the algorithms' performance, we employ a hierarchical SBM [42] to generate synthetic networks that capture nested community structures. This model extends the classical SBM by introducing multiple levels of community organization, where vertices are grouped into blocks at each hierarchical level. The parameters for the hierarchical SBM include the number of vertices (N), the number of hierarchical levels (L), a branching factor (B) where each block is divided into B child blocks, an intra-block edge probability (P_{intra}) to connect vertices within a block, and an inter-block edge probability (P_{inter}) to connect vertices between blocks.

Initially, vertices are assigned to a single root block at the topmost level (Level 0). At each subsequent level, each block is subdivided into child blocks according to the branching factor B, and edges are added based on the specified probabilities P_{intra} and P_{inter}, forming a hierarchical dendrogram structure. Additionally, since the scholarly network is weighted and its adjacency matrix represents topic-based similarity, we also assign weights to the edges in the synthetic networks. For intra-block edges, weights are randomly sampled from the range $[0.5, 1.0]$, and for inter-block edges, from the range $[0.0, 0.5]$.

The block sizes at each hierarchical level are determined by the total number of vertices, the branching factor, and the hierarchy depth. For the synthetic networks, we used $N \in \{500, 1000\}$ and $L \in \{3, 4, 5, 6\}$ with a branching factor $B = 3$, an intra-block edge probability of $P_{\text{intra}} = 0.7$, and an inter-block edge probability of $P_{\text{inter}} = 0.3$. The reason for choosing distinct intra- and inter-block probabilities is to ensure that the hierarchical structure is primarily driven by edge weights, similar to the topic-based scholarly network. As a result, we generated eight synthetic networks—four hierarchy levels for each network size (500 and 1000 vertices), to evaluate how well each algorithm captures multilevel community organization under controlled conditions.

Scholarly Network We followed a similar approach to that described in Section 4.3.2 to construct the initial networks, where LDA was used to compute topic-based similarity between researchers. Each pair of researchers was connected with an edge weighted between 0 and 1, representing their topical similarity. Using this procedure, we constructed three institutional networks for MSU, WSU, and CSU (see Chapter 3 for dataset details). In addition, we combined all three into a single aggregated network to evaluate the proposed methods on a larger, more diverse scholarly network.

As discussed in the previous chapter, connecting researchers purely based on topic similarity results in a fully connected network since similarity values are strictly greater than zero. To mitigate this, we applied edge-weight thresholds to prune weak connections and retain only meaningful relationships that reflect stronger topical alignment. In the previous chapter, these thresholds were tuned based on modularity optimization; however, higher thresholds, while increasing modularity, also risked pruning interdisciplinary connections that bridge distinct research areas. This trade-off motivated the need for hierarchical community detection explored in this chapter. By introducing hierarchy, we can analyze community structures at multiple levels of granularity and preserve potential interdisciplinary connections that may exist at higher levels of abstraction.

Although a fully connected network captures all possible relationships, it introduces significant noise and visual clutter, making it impractical for both visualization and structural interpretation. Therefore, in this chapter, we adopted a less aggressive pruning strategy—removing edges until the network reached a density of approximately 0.1. Prior work has shown that moderate network sparsity (density between 0.05 and 0.15) often facilitates clearer modular boundaries and improves the interpretability of detected communities [60, 83]. A density near 0.1 provided a balance between structural clarity and sufficient connectivity for meaningful community detection. We conducted experiments on both the fully connected and reduced-density networks to evaluate the consistency of the detected hierarchies.

The Spectral–HAC method for hierarchical community detection requires several hyperparameters to be defined, including the similarity metric, the linkage criterion for merging clusters, and the number of eigenvectors corresponding to the k smallest eigenvalues. To determine these settings, we evaluated multiple distance metrics and linkage functions using CPCC, which measures how well the resulting dendrogram preserves the pairwise similarities in the original data. Among the distance metrics tested, "cosine" similarity consistently yielded higher CPCC scores. Cosine similarity is appropriate in this setting because spectral embeddings are real-valued vectors rather than probability distributions, which is why we used Jensen–Shannon divergence earlier for topic-probability comparisons. For the linkage criterion, we examined "single", "complete", "average", and "Ward" linkage and found that "complete" linkage achieved the highest CPCC values and produced the most coherent hierarchical structures in practice.

Traditionally, in spectral clustering with k-means, the number of eigenvectors is selected either from prior knowledge of the expected number of communities or by tuning k using an evaluation metric such as modularity. However, because modularity suffers from the resolution limit problem, it is not ideal for evaluating partitions in the presence of hierarchical structures. Therefore, we used the CPCC to determine the optimal number of spectral components.

When computing CPCC for Spectral–HAC, we obtained two baseline similarity matrices: one derived directly from the topic-based adjacency matrix and another from the spectral embeddings obtained through the eigendecomposition of the normalized graph Laplacian. Figure 5.1 shows the CPCC values for both representations across $k = 2$ to 15 eigenvectors on the combined dataset. The two curves exhibit a similar overall shape, which is expected because both representations originate from the same underlying spectral structure of the graph. The Laplacian-based CPCC values are consistently higher, reflecting the fact that the hierarchical structure in Spectral–HAC is built directly from the Laplacian

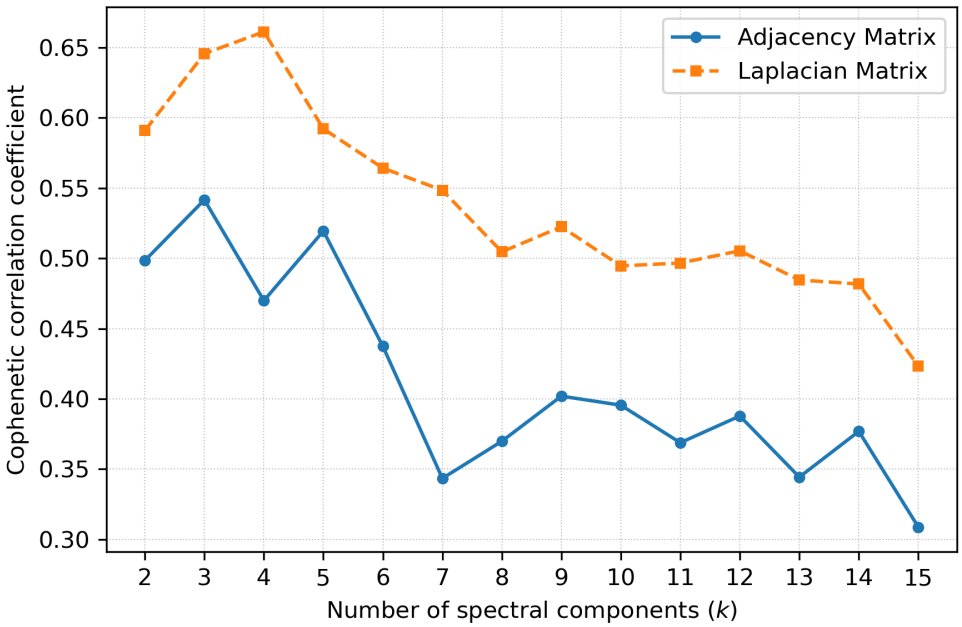


Figure 5.1: CPCC vs. spectral components for Spectral–HAC on the combined dataset

embeddings, making the cophenetic distances more closely aligned with that space. To maintain consistency with the Louvain-based approaches that operate directly on adjacency weights, we selected the number of spectral components based on the adjacency-derived CPCC values, choosing the k that achieved the highest correlation.

5.4 Result and Discussion

Table 5.1 reports the CPCC values and maximum dendrogram depths for the synthetic hierarchical networks. In the table, the columns labeled Louvain, NH-Louvain, and Spectral correspond to the base Louvain algorithm, the nested hierarchical Louvain algorithm, and Spectral–HAC, respectively. Entries labeled "A_(ℓ)" correspond to networks with 500 vertices and ℓ hierarchical levels, while "B_(ℓ)" represent networks with 1000 vertices and ℓ levels. These datasets were generated to contain explicit multi-level community structures, allowing a controlled comparison of how well each algorithm recovers the planted hierarchy.

Table 5.1: CPCC and maximum dendrogram depth for synthetic hierarchical networks

Dataset	**Louvain**		**NH-Louvain**		**Spectral**	
	CPCC	**Depth**	**CPCC**	**Depth**	**CPCC**	**Depth**
A_3	0.40	3	0.50	7	0.37	14
A_4	0.36	4	0.49	6	0.40	14
A_5	0.36	3	0.47	6	0.36	18
A_6	0.37	4	0.50	6	0.33	16
B_3	0.40	3	0.50	7	0.40	16
B_4	0.40	4	0.52	7	0.39	18
B_5	0.37	4	0.50	7	0.37	18
B_6	0.34	4	0.45	7	0.33	18

As shown in the table, NH-Louvain achieves the highest CPCC values across most configurations, confirming its robustness in hierarchically structured settings. Spectral–HAC consistently produces the deepest trees, reflecting its agglomerative nature that merges a single pair of clusters at each iteration. Louvain yields the shallowest trees and lower CPCC values, reinforcing its tendency to identify broader, coarse partitions. Collectively, these results show that NH-Louvain provides the most accurate reconstruction of hierarchical structure, while Spectral–HAC offers the greatest depth and resolution, making both methods complementary within the proposed framework.

For the synthetic networks, we also observed differences between the predefined hierarchy levels and the tree depths discovered by the algorithms. Although each synthetic network was generated with a known hierarchical structure, the addition of random edge weights introduced small deviations between the constructed hierarchy and the one recovered by each method. This deviation arises because randomized edge weights can relax the

separation between hierarchical levels, making boundaries between communities less distinct. Louvain produces maximum tree depths ranging from 3 to 4 levels, while NH-Louvain yields depths of approximately 6 to 7 levels. Since NH-Louvain recursively applies the Louvain algorithm to obtain subpartitions, it inherits a milder form of the resolution-limit effect observed in Louvain, which can constrain the number of detectable hierarchical layers. Although less pronounced than in Louvain, this limitation explains why NH-Louvain hierarchies rarely exceed seven levels. In contrast, Spectral clustering exhibits a broader range of depths, between 12 and 18 levels, owing to its pairwise merging of vertices at each iteration in the spectral embedding space. This fine-grained merging process accounts for the deeper dendrograms observed in Spectral–HAC. Together, the real and synthetic results demonstrate that NH-Louvain captures hierarchical organization most faithfully in terms of CPCC, while Spectral–HAC remains particularly effective for exploring multi-resolution community structures through highly detailed hierarchical representations.

Table 5.2 reports the CPCC values and maximum dendrogram depths obtained by the three algorithms on the real institutional networks. For each institution, we include results for the full topic-based network (MSU, WSU, CSU, and the combined MWC network) and for the pruned variant obtained by removing edges below the similarity threshold that yields a density of approximately 0.1 (MSU-0.1, WSU-0.1, CSU-0.1, and MWC-0.1).

Several patterns emerge from these results. First, for the full institutional networks (MSU, WSU, and CSU), NH-Louvain consistently achieves the highest CPCC values (0.630, 0.652, and 0.632, respectively), indicating that the recursive refinement of Louvain partitions better preserves the pairwise similarity structure than the Louvain baseline or Spectral–HAC at this level of sparsity. For the pruned networks, NH-Louvain again provides the highest CPCC for WSU-0.1, CSU-0.1, and MWC-0.1, whereas Spectral–HAC attains the best CPCC for MSU-0.1 (0.581 compared to 0.547 for NH-Louvain). This suggests that spectral embeddings can be particularly effective when weaker edges are removed and the underlying

Table 5.2: CPCC and maximum dendrogram depth for real institutional networks

Dataset	**Louvain**		**NH-Louvain**		**Spectral**	
	CPCC	**Depth**	**CPCC**	**Depth**	**CPCC**	**Depth**
MSU	0.54	4	0.63	6	0.52	14
MSU-0.1	0.50	4	0.55	7	0.58	20
WSU	0.57	3	0.65	8	0.52	16
WSU-0.1	0.55	4	0.61	8	0.58	19
CSU	0.61	4	0.63	8	0.54	16
CSU-0.1	0.57	4	0.63	8	0.61	21
MWC	0.55	4	0.61	8	0.52	18
MWC-0.1	0.49	4	0.58	8	0.53	24

communities are more sharply defined. Across all real networks, the Louvain baseline yields the lowest CPCC values, reinforcing the limitations of a single-level modularity optimization approach for capturing hierarchical structure.

The depth values in Table 5.2 also highlight important differences in the structure of the hierarchies produced by each method. Louvain consistently yields very shallow trees, with depths between 3 and 4, reflecting coarse partitions with limited hierarchical refinement. NH-Louvain produces moderately deeper trees (depths between 6 and 8), as expected from its recursive application of Louvain within each community. Spectral–HAC, in contrast, generates the deepest hierarchies, with depths ranging from 14 to 24 for the institutional networks where it was run. This behavior is consistent with its agglomerative clustering strategy, which merges only the closest pair of clusters at each step in the spectral space, resulting in a finely resolved dendrogram.

For the remainder of this analysis, we illustrate the hierarchical structure and topical

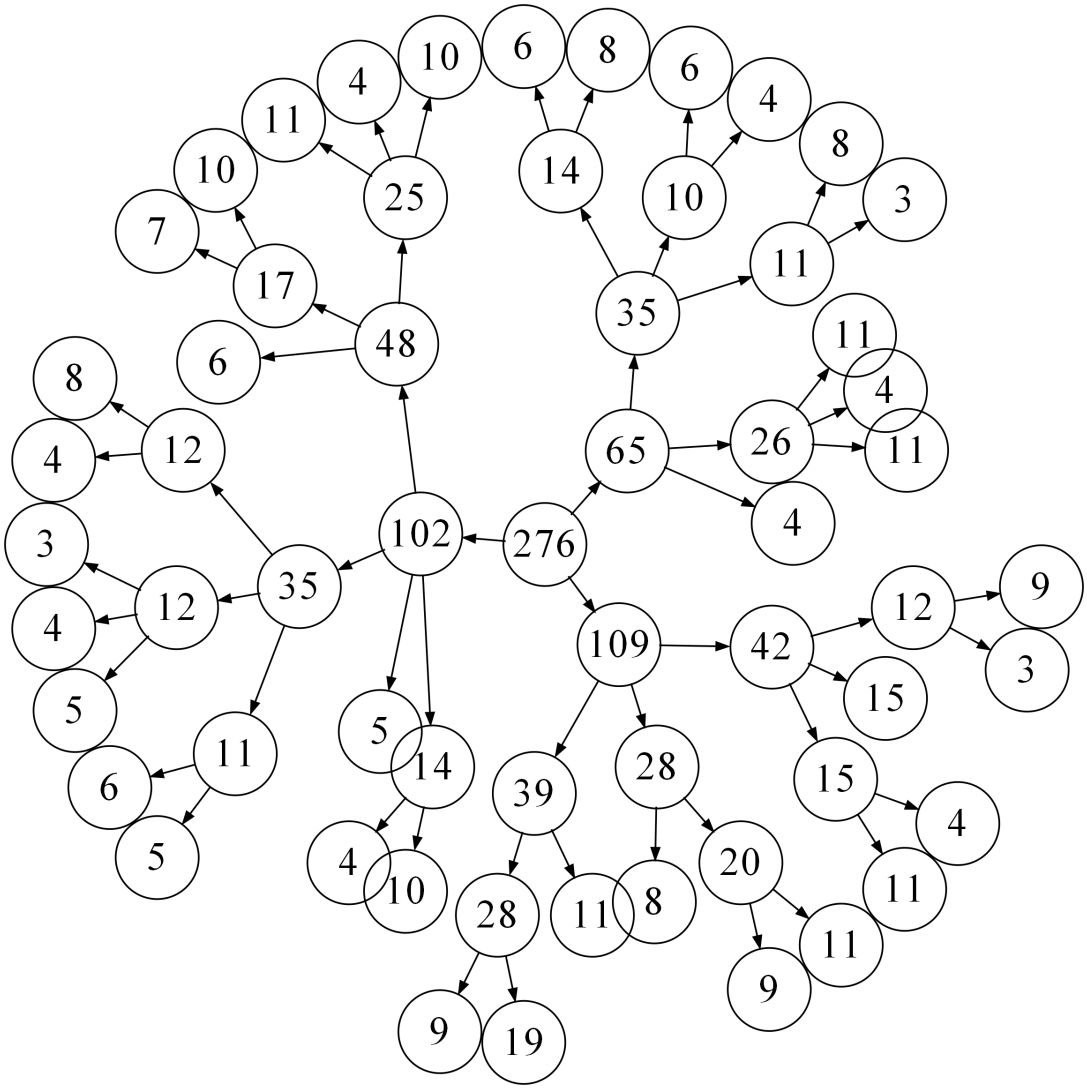
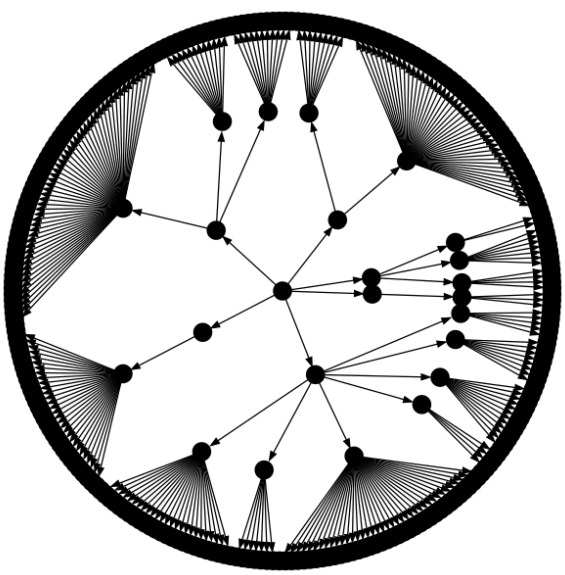


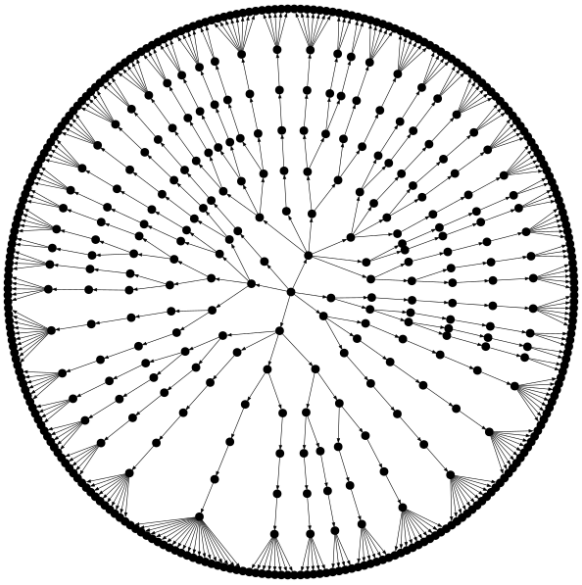
Figure 5.2: Dendrogram of the MSU network with NH-Louvain

composition using the MSU network as a representative example. The MSU network contains college and department metadata, which helps interpret the hierarchical structure more clearly than the other institutional networks. Dendrogram analyses for the WSU, CSU, and combined MWC networks are included in Appendix B.

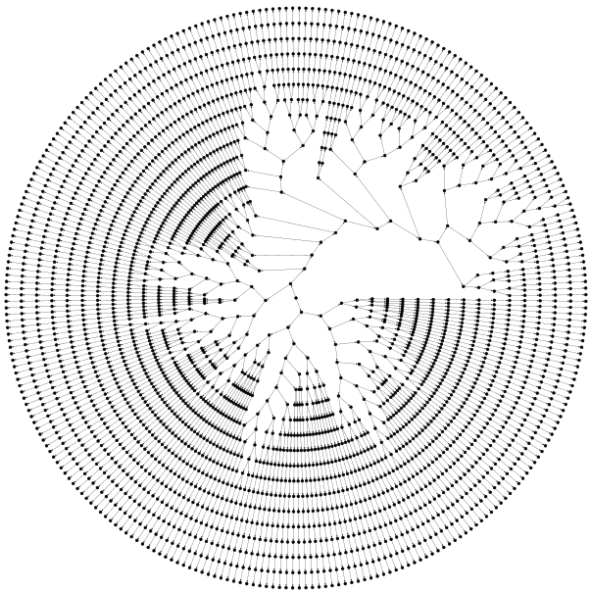
Figure 5.2 shows an example dendrogram for the MSU network generated using the NH-Louvain algorithm. The vertex values indicate the size of communities at each hierarchical



(a) Louvain



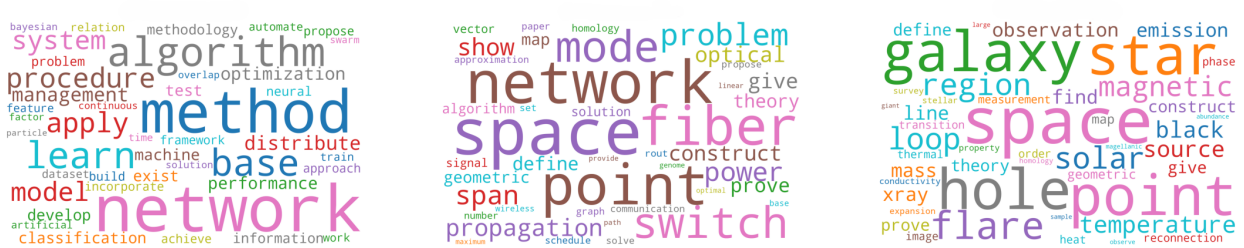
(b) NH-Louvain



(c) Spectral-HAC

Figure 5.3: Dendrograms of the MSU-0.1 network

level. The dendrogram reveals six distinct levels, approximately three levels deeper than the corresponding Louvain-based hierarchy. The overall structure is relatively full, though several



(a) WordCloud at level 4 (b) WordCloud at level 2 (c) WordCloud at level 1

Figure 5.4: Hierarchical WordCloud examples for the MSU-0.1 network

branches terminate earlier at four levels, illustrating that the hierarchical depth depends on local network structure and the density of topical connections among subsets of researchers.

Further analysis of the dendrograms for the MSU-0.1 network is shown in Figure 5.3. The base Louvain algorithm, shown in Figure 5.3a, rapidly converges to a shallow hierarchy, where the largest communities emerge within the first iteration and only limited merging occurs thereafter. This results in coarse partitions and a small number of hierarchical levels. In contrast, the NH-Louvain dendrogram in Figure 5.3b displays a more detailed and balanced structure, with recursive subdivisions producing finer community granularity. The Spectral–HAC dendrogram, shown in Figure 5.3c, extends even deeper, reflecting its agglomerative nature in the spectral space, where clusters are merged pairwise according to minimal distance. This fine-grained merging process produces dendrograms that capture subtle structural transitions between topical subcommunities. Together, these examples highlight how the proposed hierarchical models reveal varying levels of structure within the same network and enable interpretability through visualization of community hierarchies.

Figure 5.4 presents hierarchical community WordCloud examples across multiple levels of the MSU-0.1 network hierarchy. To generate these WordClouds, we aggregated publication metadata for each community and extracted representative topics using the trained topic model. At hierarchy level 4 (Figure 5.4a), the WordCloud captures fine-grained communities

primarily associated with algorithmic and machine learning topics. At higher levels, we observe broader thematic generalization: level 2 (Figure 5.4b) integrates topics such as optics and applied mathematics, while level 1 (Figure 5.4c) captures cross-domain science-related themes representing the most general layer of the hierarchy. These hierarchical WordClouds indicate that the multi-level community structures identified by NH-Louvain and Spectral–HAC correspond to semantically coherent topic groupings at different levels of abstraction.

5.5 Summary

This chapter investigated hierarchical community detection strategies to uncover the multi-level organization of scholarly collaboration networks. Building on the baseline Louvain algorithm, which provides an implicit but shallow hierarchy through modularity optimization, we developed two complementary approaches: the NH-Louvain algorithm and the Spectral–HAC algorithm. NH-Louvain extends the traditional Louvain method by recursively applying it on induced subgraphs to address the resolution limit problem and capture finer community structures. Spectral–HAC, in contrast, leverages the graph's eigenstructure with hierarchical agglomerative clustering to derive a bottom-up hierarchy grounded in the network's latent geometry. To evaluate these hierarchies, we employed the Cophenetic Correlation Coefficient (CPCC), which measures how faithfully the dendrogram preserves the pairwise similarities of the original network.

Experimental analyses across institutional and synthetic datasets provided several insights. NH-Louvain consistently demonstrated higher hierarchical fidelity than the baseline Louvain algorithm, while Spectral–HAC produced comparably deep or deeper hierarchies depending on the network's density and edge-weight distribution. Moderate edge pruning improved both interpretability and separability, indicating that retaining only mid-strength connections helps reveal the underlying multi-level organization of research

topics. Visual inspection of dendrograms and hierarchical WordClouds confirmed that the proposed methods captured meaningful transitions from specialized research themes to broader disciplinary areas across hierarchy levels.

While these findings validate the potential of hierarchical modeling for scholarly networks, several limitations remain. The evaluation in this chapter focused on structural coherence (through CPCC) without explicitly addressing semantic coherence, for example how research themes evolve across hierarchy levels. Additionally, both algorithms assume disjoint community membership, whereas real-world researchers often contribute to multiple, overlapping topical areas. Extending NH-Louvain or Spectral–HAC to support overlapping or fuzzy memberships, similar to Wahl's [122] approach, would better capture the nuanced structure of interdisciplinary collaboration.

Another key challenge concerns publication imbalance, where prolific authors disproportionately shape community structures and obscure secondary research themes. The next chapter addresses this issue by introducing strategies to balance representation and ensure fairer modeling of scholarly hierarchies.

CHAPTER SIX

HANDLING PUBLICATION IMBALANCE

In this chapter, we address the third research question of this dissertation: "How does accounting for data imbalance improve community detection and enhance collaboration recommendations in content-based social networks?" Many content-driven systems implicitly assume that all entities contribute comparable amounts of information, but real-world data rarely follow this assumption. Content can be unevenly distributed, where some entities generate large and diverse collections of items, whereas others contribute only a small amount or focus on narrow themes.

In scholarly networks, this general form of data imbalance appears specifically as publication imbalance: some researchers produce large and diverse bodies of work across multiple topics, while others publish less frequently or focus on narrower research areas. This imbalance can distort topic-based similarity scores, allowing prolific researchers to dominate the network and masking their less frequent but thematically distinct research interests.

To mitigate this issue, this chapter introduces a publication balancing strategy that explicitly models author-level heterogeneity during network construction. We propose a "cloning-based approach", where a researcher's publications are first grouped into topical clusters, allowing each cluster to represent a distinct research focus. This grouping is achieved through topic modeling using both LDA and BERTopic, where LDA offers a probabilistic and interpretable representation of topics, while BERTopic, leveraging fine-tuned SciBERT embeddings, captures contextual semantics through transformer-based representations. Each resulting cluster is then treated as a separate node in the scholarly network, enabling a single researcher to participate in multiple communities according to their different thematic interests. This approach reduces bias arising from content-data imbalance and facilitates the

discovery of more balanced and interdisciplinary collaborations.

The chapter begins with the motivation and problem statement, followed by a review of related works addressing author-level imbalance and topical diversity in scholarly networks. Next, we describe the proposed cloning methodology, detailing how topic modeling, network construction, and edge weighting are adapted for the cloned representation. We then evaluate the quality of the resulting communities empirically and demonstrate the benefits of our approach for collaboration recommendations. Finally, we discuss the broader implications of handling content-data imbalance for scholarly network analysis and conclude with observations and potential future directions.

6.1 Motivation and Problem Statement

Research collaboration plays a central role in driving scientific advancement, yet identifying meaningful new collaborations remains a complex task. While data imbalance can arise in many content-based systems, this chapter focuses on its manifestation in scholarly networks, where it appears as publication imbalance. Earlier chapters of this dissertation demonstrated that topic-based similarity can uncover connections beyond co-authorship or citation networks, offering a more inclusive view of how scholars relate through shared research interests. However, while topic-based networks provide a rich foundation for analyzing collaboration potential, they also inherit biases from the underlying publication data.

One critical source of bias arises from "publication imbalance", where researchers vary widely in their productivity. Highly prolific researchers often publish across several topics, while others may specialize in a narrower domain. When topic models such as LDA or BERTopic are used to aggregate publication content at the researcher level, this imbalance can skew the representation. Frequent authors contribute more documents to certain topics, which amplifies their influence in the similarity matrix and can overshadow their less frequent

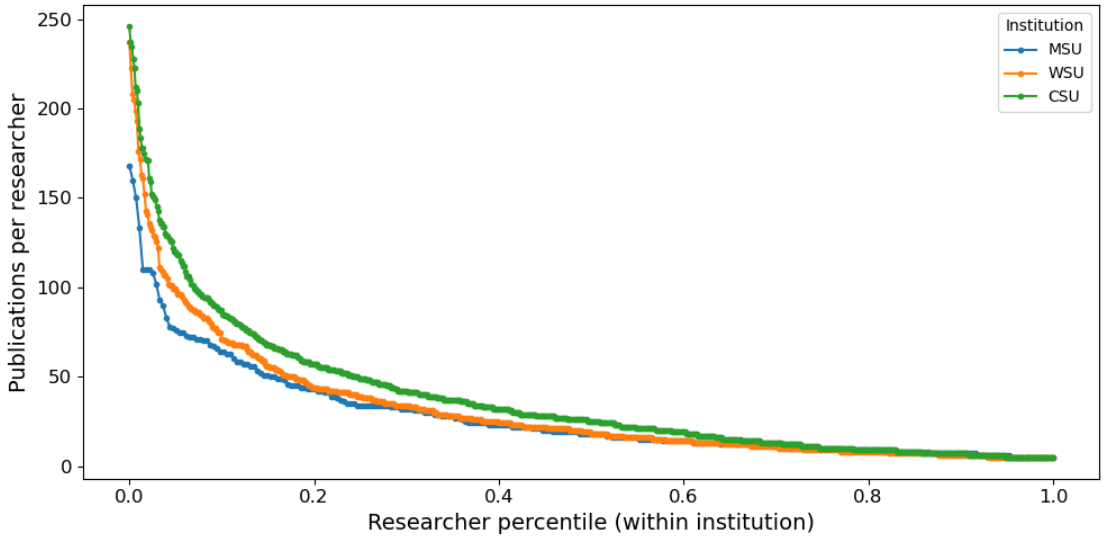


Figure 6.1: Percentile-based publication distributions across MSU, WSU, and CSU

but thematically distinct interests. Consequently, the resulting network structure tends to overemphasize dominant themes for those researchers and underrepresent emerging or secondary research areas.

This imbalance poses a problem for community detection and collaboration recommendation, particularly for prolific researchers. Without addressing this issue, their dominant theme often dictates community assignment, while secondary or diverse interests become diluted within a single representation. This limits the detection of interdisciplinary connections, as those secondary topics may serve as important bridges linking researchers across different domains. Addressing publication imbalance therefore requires a representation that preserves each researcher's diversity of interests without allowing publication volume to dominate network formation.

To illustrate the overall publication imbalance, Figure 6.1 presents the percentile-based publication distributions for MSU, WSU, and CSU. Since the total number of researchers varies across institutions, the percentile representation provides a normalized view of the publication trend. Each curve shows the number of publications per researcher against their

percentile rank within the institution, allowing a direct comparison of skewness across the three datasets. All three curves follow a similar right-skewed pattern, indicating that a small proportion of researchers contribute a disproportionately large share of the total publications. This consistent trend across institutions highlights a systemic publication imbalance within the scholarly network.

Therefore, to address this imbalance challenge, this chapter introduces a cloning-based approach that partitions each researcher's publications into topic-specific clusters, representing them as multiple nodes within the network. Each node corresponds to a distinct research focus, allowing a more balanced and context-aware structure that supports accurate community detection and diversified collaboration recommendations. This approach seeks to uncover latent relationships that are often masked by traditional author-level aggregation, providing a broader and more equitable view of scholarly collaboration potential.

Problem Statement: Given a topic-based scholarly network constructed from publication metadata, how does publication imbalance affect the resulting community structures, and how does it influence the diversity and relevance of collaboration recommendations?
Hypothesis: We hypothesize that applying a cloning-based strategy will mitigate the effects of publication imbalance by allowing prolific researchers to be represented through multiple topic-specific clusters. By capturing both dominant and secondary research areas, the cloned representation is expected to yield more balanced community structures and support more diverse collaboration recommendations.

6.2 Related Work

Earlier studies of scholarly collaboration networks primarily focused on structural relationships such as co-authorship and citation links [5, 81]. These works established fundamental properties of scientific collaboration networks but also revealed a dominance pattern where highly connected or prolific authors disproportionately shape network structures. Such structural bias often extends to content-based representations, where prolific researchers influence the formation of similarity networks more strongly than others.

Barabási and Albert [5] formalized this dominance pattern through the preferential attachment model, showing that new nodes are more likely to connect with highly connected nodes, creating a "rich-get-richer" dynamic. This effect closely resembles publication imbalance in scholarly networks, where prolific authors exert disproportionate influence on topic-based similarities and community detection outcomes. Leskovec and Faloutsos [62] also examined how large-scale graphs can suffer from structural bias and proposed sampling strategies to preserve representativeness in skewed networks. Together, these studies highlight the need to address structural dominance, which this dissertation approaches through author-level balancing rather than network resampling.

Fairness and bias mitigation have also become key considerations in recommender systems, providing a relevant perspective for scholarly collaboration modeling. Sonboli et al. [114] discussed the multi-sided nature of fairness, emphasizing that recommendation outcomes affect multiple stakeholders and must balance user and provider interests. Ekstrand et al. [26] presented a unified framework for fairness in information access systems, focusing on exposure and representational equity. Ge et al. [32] further expanded on these ideas by reviewing trustworthy recommender systems, summarizing challenges related to fairness, transparency, and robustness. Although these studies were developed in different domains, they align conceptually with the motivation of this work to ensure balanced representation

so that prolific authors do not dominate recommendation outcomes.

At the modeling level, Rosen-Zvi et al. [103] introduced the Author–Topic Model, extending Latent Dirichlet Allocation [8] to jointly infer document and author-level topic distributions. This model captures multiple topical interests for each author rather than assigning a single representation. Similarly, Griffiths et al. [34] proposed hierarchical topic models using the nested Chinese Restaurant Process to capture layered topic relationships within a corpus. While these models effectively represent topical diversity, they do not address the imbalance in publication volume across authors. In contrast, this dissertation complements such probabilistic approaches with a clustering-based cloning strategy that redistributes author representation across topic clusters.

Recent work in representation learning also supports modeling multiple perspectives of entities in networks. Zhao et al. [136] developed a self-supervised heterogeneous graph embedding framework that captures multiple relational views through multi-view consistency. Li et al. [64] and Zheng et al. [137] proposed multi-view representation learning methods that jointly optimize embeddings across complementary subspaces. Although these techniques were designed for general-purpose networks, they share a conceptual connection with this study; representing an entity through multiple embeddings helps capture context-dependent semantics. The cloning-based representation proposed in this chapter applies this principle to researchers by creating topic-specific clones that mitigate the effects of publication imbalance.

Temporal modeling has also been explored to understand evolving research interests. Jebari et al. [48] used citation contexts to analyze topic evolution over time, while Kong et al. [56] and Zhou et al. [138] incorporated temporal weighting into similarity computations to capture dynamic collaboration patterns. These studies primarily examine longitudinal changes in research focus. In contrast, the present work addresses the structural imbalance that exists within a static timeframe, focusing on the unequal contribution of prolific

researchers across concurrent topics rather than temporal evolution.

In summary, prior research has advanced our understanding of collaboration networks, topic modeling, and fairness in recommendation systems, yet none have explicitly addressed publication imbalance in topic-based scholarly networks. This dissertation extends these areas by proposing a cloning-based strategy that balances author representation, captures diverse topical interests, and improves both community interpretability and fairness in collaboration recommendations.

In summary, prior research has advanced our understanding of collaboration networks, topic modeling, and fairness in recommendation systems, yet none have explicitly addressed publication imbalance in topic-based scholarly networks. This dissertation extends these areas by proposing a cloning-based strategy that balances author representation, captures diverse topical interests, and improves both community interpretability and the diversity of collaboration recommendations. Although developed for scholarly networks, the underlying idea is more general: any content-based network in which entities contribute highly imbalanced amounts of data (such as product–item networks, user-generated content platforms, or scientific dataset repositories) could benefit from a similar cloning strategy to prevent dominant contributors from overshadowing latent or secondary themes in the data.

6.3 Methodology

Our proposed methodology, summarized in Figure 6.2, extends the topic-based scholarly network framework by addressing publication imbalance explicitly. The process begins with training topic models, namely LDA and BERTopic, on publication metadata to capture topical representations of each researcher's work. Based on the distribution of publication counts, we identify prolific researchers whose high output may bias community formation and apply a cloning-based strategy to represent their diverse topical interests.

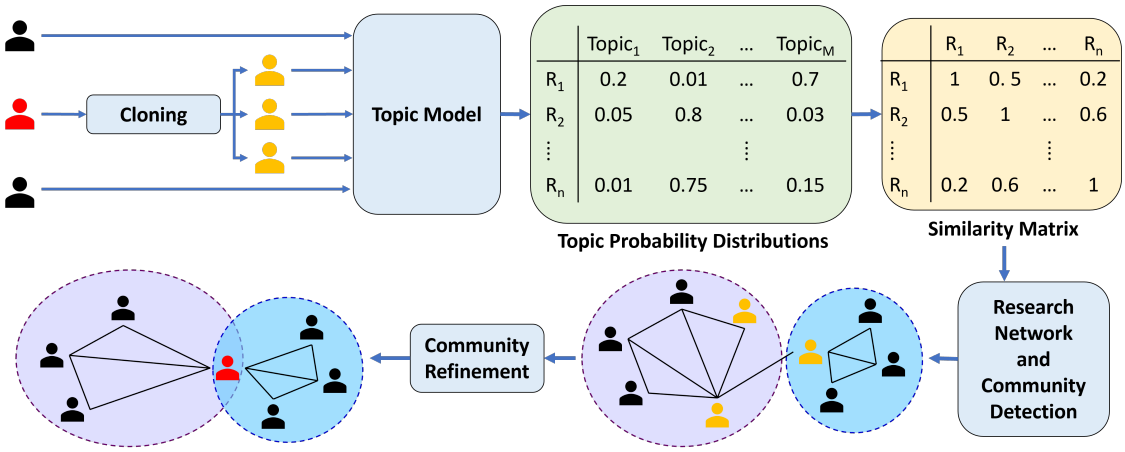


Figure 6.2: Overview of the proposed methodology

Each researcher (and their clones) is then represented through topic probability distributions, which are used to compute pairwise topic similarities and construct the research network. Community detection algorithms are applied to this network to identify groups of researchers sharing coherent topical themes. Finally, we merge cloned nodes corresponding to the same researcher to obtain the refined set of research communities that balance productivity while preserving diversity in research focus. Since cloned nodes may be assigned to different communities, this process also allows a single researcher to have overlapping community memberships, reflecting the multi-faceted nature of their research interests. Although we do not employ dedicated overlapping community detection algorithms, cloning provides a natural mechanism for capturing such overlap, and comparing this approach with formal overlapping methods remains a potential direction for future work

6.3.1 Topic Modeling

Topic modeling serves as the foundation for constructing topic-based scholarly networks, where topical similarity between researchers is derived from the distribution of research themes across their publications. In this study, we employ two topic modeling approaches: LDA and BERTopic. LDA provides a probabilistic and interpretable baseline, while

BERTopic introduces contextual embedding-based representations capable of capturing semantic relationships beyond word co-occurrence. Although the primary objective of this chapter is not to compare the performance of topic models, using both methods offers complementary perspectives for building topic-based researcher networks and evaluating the effects of publication imbalance.

LDA As introduced in Section 4.3.1, LDA [8] models each document as a mixture of latent topics and each topic as a distribution over words. This generative framework assumes that documents are produced by first sampling a distribution of topics, then sampling words conditioned on those topics. LDA remains one of the most widely adopted probabilistic models for uncovering latent thematic structures in text due to its interpretability and simplicity.

In this chapter, LDA is employed to extract document-level topic distributions from publication titles and abstracts. Each researcher's topic profile is computed by concatenating all their publications into a single composite document and extracting topic probabilities from the trained LDA model. The resulting topic mixture vector reflects the researcher's overall topical composition and serves as input for computing researcher-level topical similarity and constructing the topic-based scholarly network used in subsequent analyses. The motivation for the cloning strategy also stems from this representation: concatenating the publications of prolific researchers, sometimes exceeding 150 documents, into a single large document may cause dominant topics to overshadow their secondary or interdisciplinary interests.

BERTopic While LDA provides interpretable topic mixtures, its bag-of-words representation limits its ability to capture semantic and contextual relationships. To address these limitations, we employ BERTopic [35], a transformer-based topic modeling framework that integrates deep contextual embeddings with dimensionality reduction, clustering, and keyword extraction. With the rapid advancement of transformer-based models, BERTopic

has become a popular and effective approach for discovering semantically coherent topics from unstructured text. The BERTopic pipeline consists of the following main components:

- **Sentence Embedding:** Each document is represented using contextual embeddings generated by a pre-trained transformer model [7, 27, 98]. These embeddings capture semantic relationships between words and phrases by considering their surrounding context, allowing conceptually similar documents to occupy nearby positions in the embedding space.

- **Dimensionality Reduction (UMAP):** Since transformer-based embeddings are high-dimensional, Uniform Manifold Approximation and Projection (UMAP) [78] is applied to project the document embeddings into a lower-dimensional space. This step preserves both local and global structure, helping reveal the intrinsic geometry on which documents lie and facilitating clustering.

- **Clustering (HDBSCAN):** The reduced embeddings are clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [77]. HDBSCAN identifies clusters of varying density, handles noise points, and does not require specifying the number of clusters, making it well suited for scholarly datasets in which topic density varies across fields. Although BERTopic is commonly paired with HDBSCAN, the framework is modular and supports other clustering algorithms (e.g., k-means, spectral clustering, or agglomerative clustering), with HDBSCAN chosen here for its ability to adapt to irregular cluster shapes and differing density levels.

- **Topic Representation (c-TFIDF):** For each cluster, class-based term frequency–inverse document frequency (c-TFIDF) [35] is applied to extract representative keywords. Unlike standard TF-IDF, which treats each document independently, c-TFIDF aggregates all documents in a cluster into a single class document. This highlights

terms that distinguish one cluster from the rest of the corpus and is essential for producing interpretable topic descriptors in BERTopic. The resulting term weights provide a pseudo-probability over words, analogous to the topic-word distributions produced by LDA.

In this work, BERTopic is implemented with SciBERT [7] as the embedding model, as SciBERT is pre-trained on a large corpus of scientific text and is thus well suited for capturing the linguistic and conceptual nuances of scholarly publications. To further improve topic quality and alignment, SciBERT is fine-tuned on our corpus using a masked language modeling (MLM) [23] objective before generating embeddings. During fine-tuning, a random subset of tokens (typically 15%) is replaced with a special [MASK] token, and the model learns to predict the masked words based on their surrounding context. This process adjusts the pre-trained model's parameters to better represent domain-specific terminology and writing patterns in our dataset. Prior research has shown that fine-tuning transformer-based models on task-specific or domain-specific corpora can substantially improve their representational performance [25, 37, 119].

Although SciBERT is well matched to scientific text, we note that fine-tuning on a domain-specific corpus may introduce new biases by reinforcing terminology and stylistic patterns that are more common in the sampled institutions or disciplines. This limitation is not unique to SciBERT, as any domain-adapted model may underrepresent less frequent or non-scientific linguistic structures, and should be considered when interpreting topic assignments. Nevertheless, in the context of this work, the benefits of improved topical alignment outweigh these limitations, as the goal is to model research themes within scholarly publications.

Fine-tuning allows SciBERT to better adapt to the vocabulary and contextual dependencies unique to our institutional publication corpus, resulting in more coherent topic clusters and more accurate similarity estimation between documents. As shown in

the results section, this fine-tuned variant yields higher topical coherence and improved researcher similarity scores in the constructed networks.

For each researcher, BERTopic produces document-level topic memberships that are aggregated to form researcher-level topic vectors analogous to the LDA-based topic mixtures. These vectors are subsequently used to compute pairwise topical similarity between researchers, forming the weighted edges of the topic-based scholarly network used in downstream community detection and imbalance-handling processes.

6.3.2 Cloning Prolific Researchers

This section describes how prolific researchers are identified and how the cloning-based strategy is applied to address publication imbalance across the three institutions in this study—MSU, WSU, and CSU (see Chapter 3 for dataset details). While the core concept remains the same, creating multiple representations or "clones" for prolific researchers by clustering their publications into distinct topical themes, we employ two complementary approaches that differ in their underlying modeling framework.

The first is an LDA-based cloning approach, where a local LDA model is trained for each prolific researcher. The resulting document–topic probability distributions are then clustered using HDBSCAN to group documents sharing similar topical patterns. HDBSCAN is used here because it can identify clusters of varying density and does not require specifying the number of clusters in advance, although in principle any clustering algorithm could be employed for this step. The second approach is an embedding-based method that uses fine-tuned SciBERT to generate contextual sentence embeddings for each publication. These embeddings are clustered using the same HDBSCAN setup as in the LDA-based approach, yielding groups of semantically related documents that represent a researcher's distinct topical focuses.

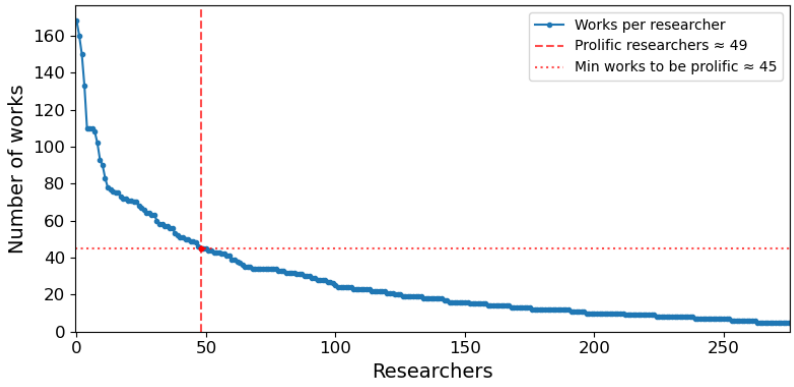
The motivation for adopting both approaches lies in their complementary strengths:

LDA provides interpretable, probabilistic topic structures, whereas transformer-based embeddings capture nuanced semantic relationships but lack direct interpretability. Together, they offer two perspectives on representing researcher diversity. After obtaining topical clusters for each prolific researcher, each cluster is treated as a separate entity—or "clone"—in the network construction process, allowing the resulting topic-based network to reflect both depth and diversity in scholarly focus.

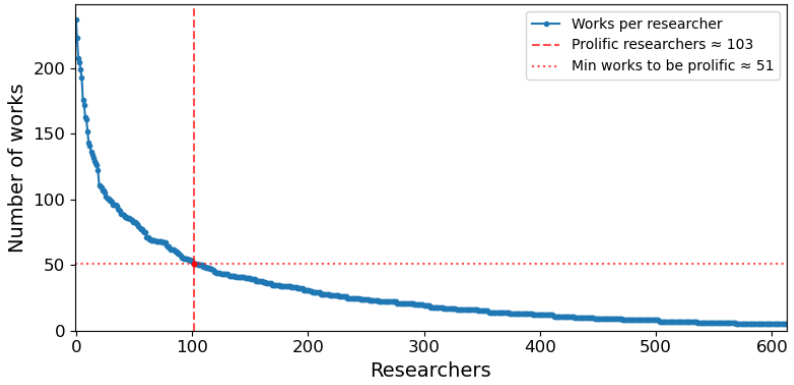
Filtering of Prolific Researchers Data (publication) imbalance is a key challenge in topic-based network construction, as a small subset of researchers often contributes a disproportionately large share of publications. This imbalance can bias community detection, since prolific researchers may dominate topical similarity measures and influence network structure. To quantify this imbalance and identify prolific researchers, we first examined the publication count distribution across the three institutions in this study (MSU, WSU, and CSU).

As shown in Figure 6.3, all three institutions show a pronounced right-skewed distribution, indicating that a small number of researchers publish substantially more than the rest. This detailed view complements the combined imbalance pattern presented earlier in Figure 6.1, providing the institution-specific trends used in the cloning methodology. To determine which researchers qualify as prolific, we applied the "elbow method" [14] to the ranked publication counts. The elbow method is a heuristic commonly used to identify a threshold or cutoff point in a curve where the rate of change sharply decreases, resembling an "elbow" shape. In this context, it detects the inflection point in the publication count curve where adding more researchers yields diminishing returns in total output. Researchers with publication counts beyond this elbow point are considered prolific.

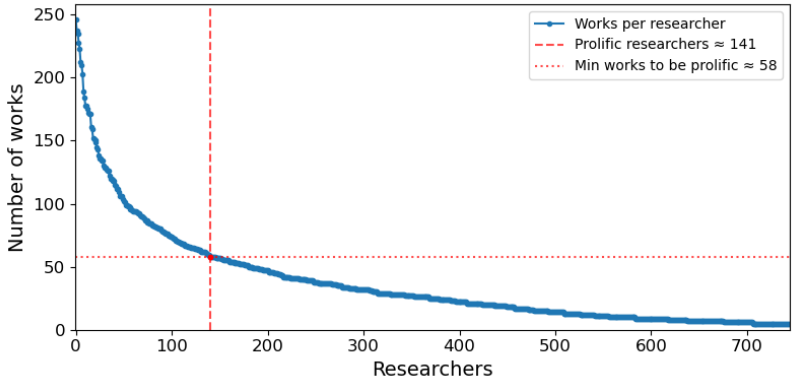
Applying this approach, the prolific publication thresholds were determined to be 45 for MSU, 51 for WSU, and 58 for CSU, respectively. These thresholds define the subset



(a) MSU



(b) WSU



(c) CSU

Figure 6.3: Distribution of researcher publication counts across MSU, WSU, and CSU.

of researchers for whom the cloning strategy, described in the next sections, is applied to mitigate the dominance of publication volume in community detection.

LDA-based Cloning For the LDA-based cloning approach, we begin with the set of publications authored by each prolific researcher and cluster them based on topical similarity. To derive document-level topic representations, two modeling strategies can be considered: a global LDA model trained on the entire corpus or a local LDA model trained on each researcher's publications.

In the global approach, an LDA model is trained using all available publications, and the resulting topic distributions are used to represent individual documents. However, this approach presents a key limitation. The global model often contains a large number of topics (e.g., 100–150), while the topical scope of a single researcher's publications is typically much narrower. As a result, the document–topic vectors become highly sparse, with non-informative probabilities for irrelevant topics. This sparsity can reduce the model's ability to cluster publications at the researcher level meaningfully.

To address this issue, we employ a local LDA approach, training a separate LDA model for each prolific researcher. The number of topics in these local models is considerably smaller than in the global model, making the resulting topic distributions more representative of each researcher's actual thematic range. This produces denser document–topic vectors that more accurately reflect the topical composition of a given researcher's work and provide a better input space for clustering. As discussed later in the experimental design section, different topic ranges were tested during local training to ensure stability of the results.

Once the local LDA model is trained for a prolific researcher, we obtain a topic-probability vector for each publication, which serves as the feature representation for clustering. The HDBSCAN algorithm is then applied to these vectors to group publications according to shared topical patterns.

A potential concern with this approach is scalability—training an independent LDA model for every prolific researcher may appear computationally expensive for larger datasets. In practice, this step remains feasible because the number of prolific researchers is relatively small, and each researcher typically has between 45 and 300 publications. Since each model is trained independently, the process can be parallelized easily. Moreover, LDA with collapsed Gibbs sampling converges efficiently for such dataset sizes. If scalability becomes a constraint in larger settings, a global LDA model could still be used with dimensionality reduction or topic filtering techniques to reduce sparsity, though these alternatives may provide less researcher-specific detail.

After clustering the topical representations with HDBSCAN, each resulting cluster is treated as a separate clone for that researcher. For example, if a researcher's publications are grouped into three clusters, three clones—A1, A2, and A3—are created, each corresponding to one distinct topical cluster. HDBSCAN may also label certain publications as noise (assigned the label -1); since these documents may still capture secondary or emerging themes, they are retained as an additional clone to preserve a complete view of the researcher's topical diversity.

BERT-based Cloning The BERT-based cloning approach follows a similar overall procedure to the LDA-based method, but relies on transformer-based sentence embeddings rather than probabilistic topic distributions to represent documents. In this approach, each publication is first converted into a dense numerical vector using a sentence embedding model, which captures contextual relationships between words and phrases within the text.

We employ SciBERT [7] as the base embedding model. As discussed earlier in the BERTopic description, SciBERT is pre-trained on scientific and technical text, providing good initial coverage of the vocabulary and stylistic patterns commonly found in scholarly publications. To better align the model with our specific corpus, we fine-tune SciBERT

using a masked language modeling (MLM) objective on all publications across the three institutions. In MLM, a portion of tokens is replaced with a [MASK] symbol, and the model learns to predict the missing words from surrounding context. This process encourages the model to internalize corpus-specific terminology and usage patterns, resulting in document embeddings that more accurately reflect the semantic structure of our dataset.

Each publication is then represented by the embedding vector generated by the fine-tuned SciBERT model. These embeddings are high-dimensional, which can hinder efficient clustering. To address this, we apply UMAP to reduce the embedding space while preserving both local and global semantic relationships. After testing multiple dimensionality settings, we fixed the reduced dimensionality to five for all experiments, as detailed later in the experimental design section.

The reduced document embeddings are then clustered using HDBSCAN to group similar publications for each prolific researcher. Each resulting cluster represents a distinct topical theme and is treated as a separate clone, consistent with the LDA-based cloning procedure. For instance, if a researcher's publications form three clusters, three clones (e.g., B1, B2, and B3) are created, each corresponding to a different area of research focus. As before, any documents labeled as noise (-1) by HDBSCAN are retained as an additional clone, since these often reflect marginal or emerging topics that still contribute to the researcher's overall research profile.

6.3.3 Research Networks with Clones

Following the cloning step, which expands the total number of researchers by introducing cloned nodes for prolific individuals, the next stage involves constructing the topic-based research network. This begins with computing topic probability distributions for all researchers, including their clones, using both global LDA and BERTopic models trained on the full corpus. The resulting distributions are then used to compute topic similarity

scores between researchers, forming the adjacency matrix of the scholarly network.

For LDA, the process closely follows the network construction described in Section 4.3.2. Each researcher's topic distribution is obtained by concatenating all their publications into a single large document and querying the trained global LDA model for the resulting topic mixture. To measure topical similarity between researcher pairs, we use the Jensen–Shannon Divergence (JSD) [104], a symmetric variant of the Kullback–Leibler divergence. The JSD produces bounded similarity scores between 0 and 1, where higher values indicate stronger topical alignment. These scores define the weighted edges of the network.

In the case of BERTopic, the fine-tuned model does not directly yield probabilistic topic distributions as LDA does. Instead, it estimates a pseudo–topic probability distribution for each document. Specifically, BERTopic divides each document into overlapping segments (token windows), assigns topics to each segment, and aggregates the assignments to estimate a normalized topic vector. To compute researcher-level topic distributions, we aggregate and normalize the topic vectors of all publications authored by a given researcher. Let $\mathbf{N}$ denote the full set of publications and $\mathbf{N}_i \subseteq \mathbf{N}$ represent the publications of researcher i with $n_i = |\mathbf{N}_i|$. If $\boldsymbol{\theta}_{p_j}$ is the topic distribution of publication $p_j \in \mathbf{N}_i$, then the researcher-level topic distribution $\boldsymbol{\Theta}_i$ is defined as:

$$\boldsymbol{\Theta}_i = \frac{1}{n_i} \sum_{p_j \in \mathbf{N}_i} \boldsymbol{\theta}_{p_j}$$

The pairwise JSD between these researcher-level topic distributions is then used to construct the weighted topic similarity matrix for BERTopic. Similar to the LDA-based network, this produces a fully connected graph. However, as discussed in the previous chapter, weak edges are pruned to achieve an optimal network density, improving the interpretability of community detection results. The difference in this setting is that the cloned researchers now appear as separate nodes, each corresponding to a distinct topical

cluster derived from the cloning process.

6.3.4 Community Detection and Refinement

In Chapter 4, we employed flat community detection algorithms such as Louvain and Spectral clustering to detect communities at a single level of granularity. Later, in Chapter 5, we demonstrated that research topics often exhibit hierarchical structures, and that a hierarchical community detection framework can better capture multi-level patterns of scholarly organization. Building on those findings, this chapter applies the Nested Hierarchical Louvain (NH-Louvain) algorithm (Section 5.3) to detect communities within the cloned research network.

Briefly, NH-Louvain begins with the standard Louvain method [9], which greedily optimizes modularity to identify communities. It then recursively applies the Louvain process to each detected community, producing a hierarchy of clusters until a stopping criterion is reached. The stopping criterion determines the desired level of granularity in the resulting hierarchy. For this study, we set a minimum community size threshold as the stopping criterion, favoring moderately sized communities that reflect broader interdisciplinary alignments rather than narrowly focused research groups. This configuration supports our goal of identifying diverse and cross-disciplinary collaboration opportunities.

Introducing clones for prolific researchers allows them to appear multiple times in the network, reflecting their engagement across different topical areas. Consequently, some clones of the same researcher may occur within the same community, while others may appear across different communities. To ensure accurate representation, we perform a community refinement step to merge all clones that appear within the same community, while retaining those that occur in distinct communities as separate entities. This refinement is conducted independently within each community subgraph rather than across the entire network. The rationale is that if a researcher's clones belong to different communities, this separation

Algorithm 6.5 Community Refinement

Require: Community subgraph G_C, community nodes C
Group nodes in C by base researcher identity
Initialize new community graph G'
for all node u in C **do**
    Add base identity of u to G'
    for all neighbor v of u in G_C **do**
        Get base identity of v
        if base identities of u and v differ **then**
            Add base identity of v to G' if not present
            Add or update edge (u, v) in G' with maximum weight
return G'

reflects genuine participation in distinct research themes and should be preserved. The refinement process is outlined in Algorithm 6.5.

The algorithm takes as input the subgraph corresponding to a detected community and returns a refined version where all clones sharing the same base identity are merged into a single node. For each node u, its base identity is added to the refined graph, and edges are aggregated between distinct base identities using the maximum observed edge weight. The resulting community graph retains all unique researcher connections while removing redundancy introduced by cloned nodes.

6.3.5 Experimental Design

This section outlines the design choices and hyperparameter settings used for training the topic models, identifying and cloning prolific researchers, and constructing the scholarly networks. Four datasets were used in total: three institutional datasets (MSU, WSU, and CSU; see Section 3 for details) and one combined dataset, denoted as MWC, which integrates all three institutions. The combined MWC dataset includes approximately 1,700 researchers and 47,000 publication records.

For the global LDA model, the training process follows the procedure described in Section 4.3.1. Briefly, the number of topics is a predefined hyperparameter that determines

the model's granularity. For the institutional datasets, we trained models with topic counts ranging from 50 to 150, in increments of 10. For the larger MWC dataset, the topic range was extended from 100 to 200, also in increments of 10. The optimal number of topics for each dataset was selected based on the `C_V` coherence measure (see Section 2.2.4), resulting in 80, 80, 90, and 160 topics for MSU, WSU, CSU, and MWC, respectively.

For BERTopic, we used Hugging Face's pre-trained allenai/scibert_scivocab_uncased[1] model as the base sentence transformer. To better align with the scholarly domain of our corpus, the model was fine-tuned using the masked language modeling (MLM) objective over all publication titles and abstracts in the MWC dataset. During fine-tuning, 15% of tokens were masked at random, and the model was trained for 20 epochs while recording the `C_V` coherence score at each epoch to identify the optimal checkpoint for downstream BERTopic training. Figure 6.4 shows the `C_V` coherence scores for BERTopic models trained using embeddings from each fine-tuning epoch of SciBERT. Based on these results, epoch 14 was selected as the optimal checkpoint for embedding generation in subsequent experiments.

For dimensionality reduction, UMAP was used to project document embeddings into a lower-dimensional space prior to clustering. UMAP was configured with n_neighbors=15, min_dist=0.1, and n_components=5. The reduced embeddings were then clustered using HDBSCAN, configured with min_cluster_size=15, min_samples=15, and cluster_selection_method='eom'.

It is important to distinguish between text representation and model training. During BERTopic training, raw (unprocessed) text was used to preserve contextual relationships between words, which are critical for transformer-based embeddings. However, for the topic representation stage, where keywords are extracted using c-TFIDF, we applied the same text preprocessing steps described in Section 3.2.2, including lowercasing, stopword removal, and lemmatization. This ensures that topic word lists remain interpretable while embeddings

[1] `https://huggingface.co/allenai/scibert_scivocab_uncased`

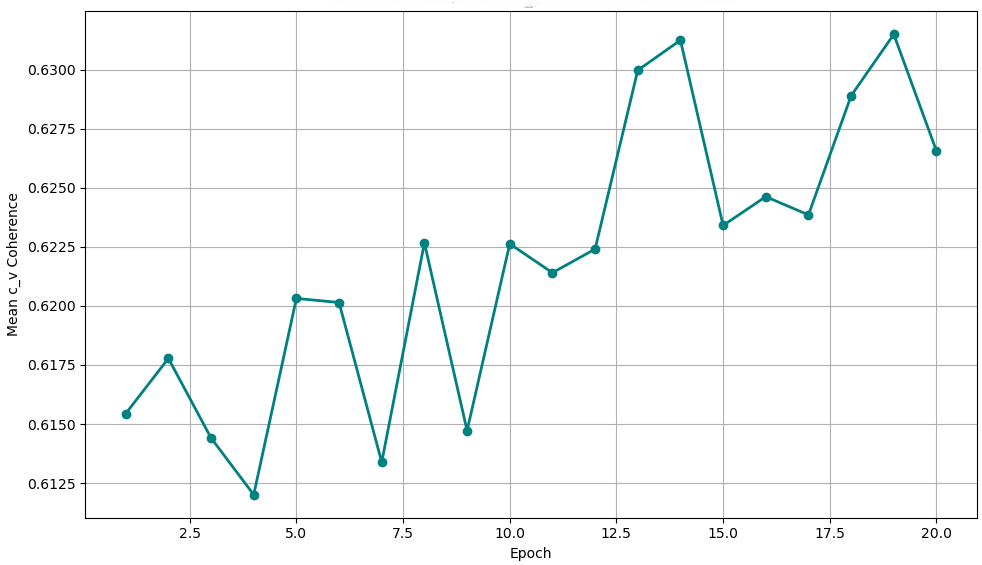


Figure 6.4: BERTopic coherence across SciBERT fine-tuning epochs on the MWC dataset

retain full contextual information. After training BERTopic on the four datasets, the number of discovered topics was 80 for MSU, 231 for WSU, 292 for CSU, and 510 for MWC. These results illustrate that BERTopic, being context-aware, tends to produce finer-grained topical structures than probabilistic LDA.

For LDA-based cloning, we trained a local LDA model for each prolific researcher using their individual set of publications (title and abstract). Each model was trained with the number of topics ranging from 2 to 10, in increments of 1, based on the intuition that a researcher-specific topic model should exhibit a narrower topical focus compared to the global model. Similar to the global LDA, we used the `C_V` coherence score to select the optimal number of topics for each researcher. The resulting document–topic probability distributions were then used as the feature space for the HDBSCAN algorithm to cluster publications according to topical similarity.

Since the topic dimensionality did not exceed ten, no additional dimensionality reduction was applied. HDBSCAN was configured with min_cluster_size=5 and min_samples=5,

Table 6.1: Cloning statistics by institution for LDA and BERT

Institution (Prolific)	**Method**	**Max**	**Min**	**Median**	**Std**	**Single-Clone**
MSU (48)	LDA	6	1	3	1.41	5
	BERT	11	1	4	1.88	2
WSU (102)	LDA	12	1	3	1.80	12
	BERT	11	1	4	2.00	3
CSU (140)	LDA	15	1	3	1.93	14
	BERT	13	1	4	2.26	3

where min_cluster_size defines the minimum number of points required to form a cluster, and min_samples specifies the number of neighboring points that determine whether a point is considered a core point. Outlier publications identified by HDBSCAN were grouped together as a separate clone when their count exceeded five.

The BERT-based cloning process was computationally simpler than the LDA approach. For each prolific researcher, we first generated document embeddings using the fine-tuned SciBERT model. These embeddings served as the feature space for clustering. Given that SciBERT produces 768-dimensional embeddings, dimensionality reduction was first performed using UMAP with n_neighbors=15, min_dist=0.0, and n_components=5. The reduced embeddings were then clustered using HDBSCAN with the same configuration as the LDA-based approach (min_cluster_size=5, min_samples=5).

Table 6.1 summarizes the cloning statistics for both LDA- and BERTopic-based methods across the three institutions. For both methods, the "Min" value is 1, indicating that some prolific researchers, despite meeting the selection threshold, were not partitioned into multiple topical clusters. This typically occurs when all of a researcher's publications are highly similar or when HDBSCAN fails to form a cluster and instead classifies all documents

as outliers. The "Single-Clone" column represents the number of prolific researchers who produced exactly one clone. Overall, the LDA-based approach yielded a higher number of single-clone cases, while BERTopic tended to generate more clones per researcher, reflecting methodological differences in how each approach captures topical diversity within an individual's publication set.

We employed the NH-Louvain algorithm for hierarchical community detection. To reduce network density and obtain more interpretable structures, edges were pruned until the overall density reached 0.1, consistent with the approach described in Chapter 5. To promote interdisciplinary community formation, a minimum community size of 30 was used as the stopping criterion, meaning NH-Louvain stopped the recursive partitioning process once a community size reached 30 or fewer members. All hyperparameters were tuned through random search, aiming for stable and interpretable community structures that reflected meaningful topical groupings.

All experiments were conducted on the Tempest High Performance Computing System, operated and supported by University Information Technology Research Cyberinfrastructure (RRID:SCR_026229) at Montana State University. Each job utilized one compute node with 32 CPU cores, one NVIDIA A40 GPU, and 64 GB of system memory. All experiments were performed separately on the three institutional datasets (MSU, WSU, and CSU) as well as the combined MWC dataset, maintaining consistent parameter settings across runs.

6.4 Result and Discussion

We begin our results discussion by examining whether cloning improves community detection among researchers, particularly those with diverse or high-impact publication records. The underlying hypothesis is that the topic distributions of prolific researchers may be underrepresented in the original network, as their dominant topical areas can overshadow other research interests. By cloning these researchers through clustering their

Table 6.2: Number of vertices in the pre- and post-cloned networks across institutions

Institution	**Pre-Cloned**	**Post-Cloned (LDA)**	**Post-Cloned (BERT)**
MSU	296	395	415
WSU	613	878	977
CSU	745	1112	1288
MWC	1634	2385	2544

publications, we aim to reveal more accurate and potentially overlapping communities that better capture interdisciplinary connections and support more diverse collaboration recommendations. Although our approach is unsupervised and lacks a ground truth for direct validation, we present both quantitative and qualitative analyses to assess its impact.

We first analyze the shift in edge-weight distributions for the pre- and post-cloned networks using both LDA- and BERT-based cloning strategies. Here, each edge weight represents the topical similarity between a pair of researchers, and our goal is to determine whether cloning increases the overall topical similarity across the network. Before examining the distributions, Table 6.2 summarizes the change in the number of vertices (researchers) before and after cloning for each institution.

As shown in Table 6.2, the number of vertices differs between the pre- and post-cloned networks. To enable a fair comparison of edge-weight distributions, we scale both networks to the same number of nodes and edges. This is achieved using the community refinement step described in Section 6.3.4, applied to the post-cloned network while treating it as a single large community. During refinement, all cloned nodes corresponding to the same researcher are merged, and when multiple edges exist between the same pair of researchers, the maximum edge weight is retained to preserve the strongest topical similarity. This process restores a network equivalent in size to the pre-cloned version, allowing a direct

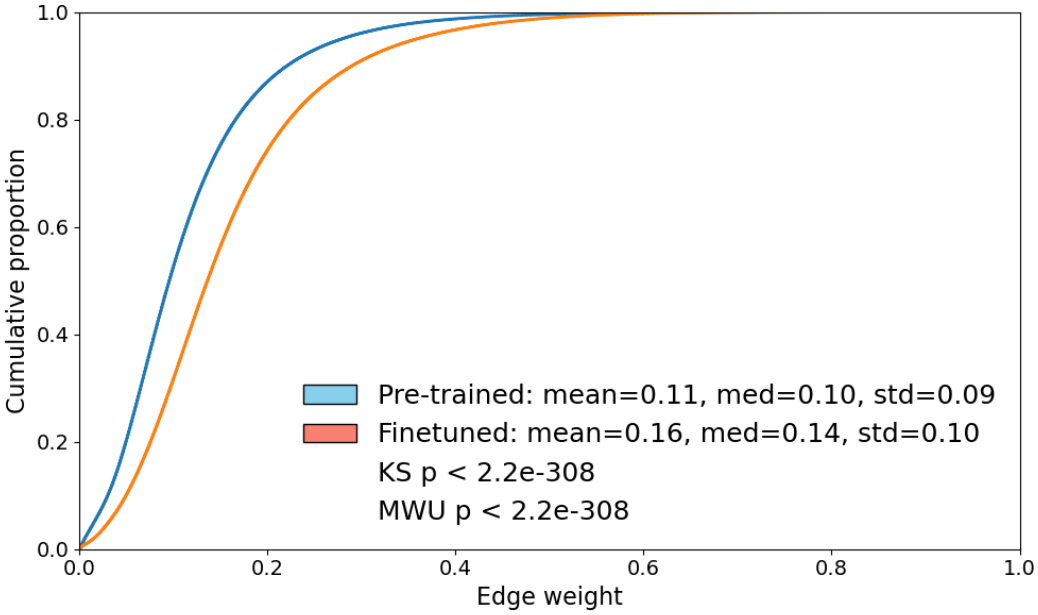
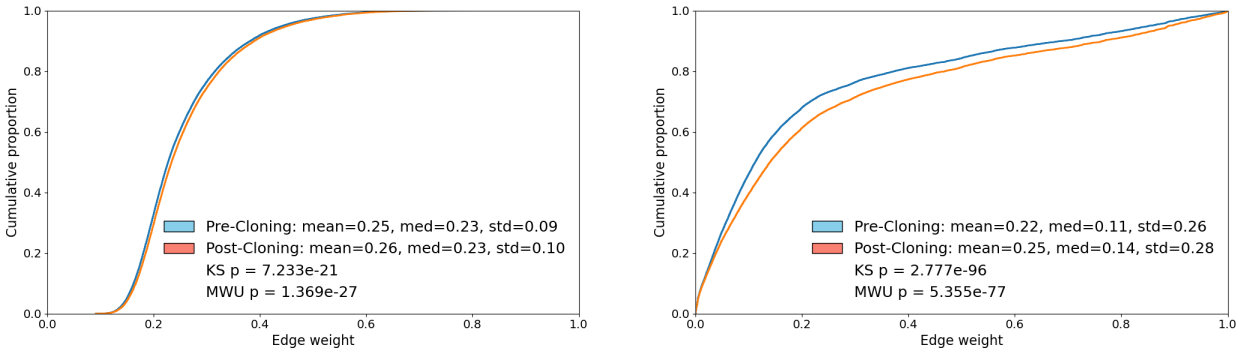


Figure 6.5: Comparison of edge-weight percentile distributions for the combined MWC network using pre-trained and fine-tuned SciBERT models

comparison of the edge-weight distributions.

Before comparing LDA- and BERT-based cloning across institutions, we first evaluate the effect of fine-tuning SciBERT on topic quality in the non-cloned version of the network. Figure 6.5 compares the pre-trained and fine-tuned versions of SciBERT on the combined MWC network. Here, the X-axis represents the edge weights, and the Y-axis represents the proportion of all edges with weights less than or equal to a given value. The figure also reports the edge statistics, including the mean, median, and standard deviation, for both distributions.

As shown, all three metrics exhibit higher values in the fine-tuned model, indicating stronger topical relationships between researchers. This difference is also statistically significant: both the Kolmogorov–Smirnov (KS) test [76] and the Mann–Whitney U (MWU) test [75] yield extremely small p-values, confirming that the two edge-weight distributions are not drawn from the same underlying population.



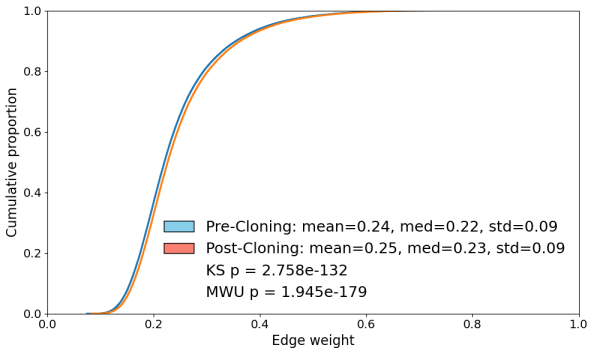
(a) MSU (LDA) (b) MSU (BERT)

Figure 6.6: Comparison of pre- and post-cloning edge-weight distributions for MSU using LDA- and BERT-based cloning

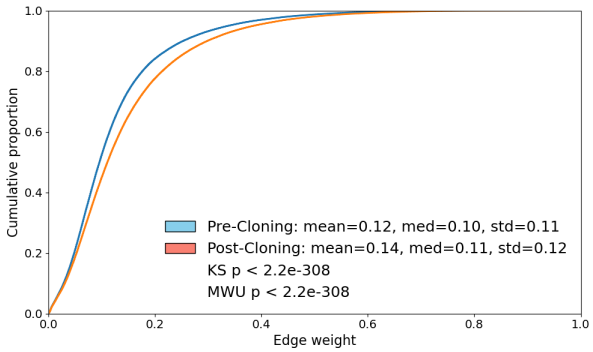
Fine-tuning further produces a clear rightward shift in the edge-weight distribution, even before applying cloning. This behavior arises because fine-tuning adapts SciBERT to the specific vocabulary, research themes, and co-occurrence patterns present in the MWC corpus. As a result, document embeddings become more domain-aligned, producing sharper and more coherent BERTopic clusters. When researcher-level embeddings are aggregated from these improved document representations, authors working on related topics are placed closer together in embedding space, yielding higher edge weights. All subsequent BERT-based analyses in this section therefore use the fine-tuned SciBERT model.

We next compare the LDA- and BERT-based cloning results across institutions. Each of the following figures shows two horizontal subplots, with the left panel representing the LDA-based edge-weight distribution and the right panel showing the BERT-based distribution using fine-tuned SciBERT.

Figures 6.6–6.8 compare the edge-weight distributions before and after cloning for each institution using both LDA- and BERT-based methods. Each subfigure presents two curves representing the pre- and post-cloned networks, where the X-axis shows the edge weights and the Y-axis shows the cumulative proportion of edges with weights less than or equal to a given

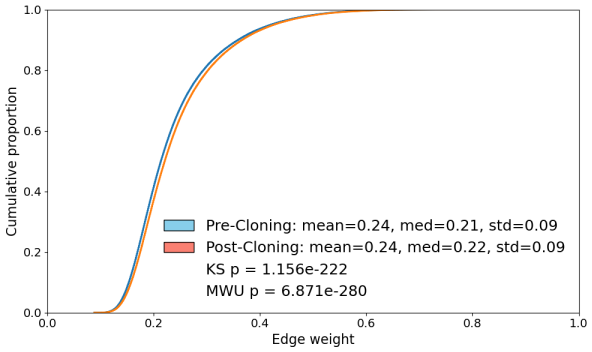


(a) WSU (LDA)

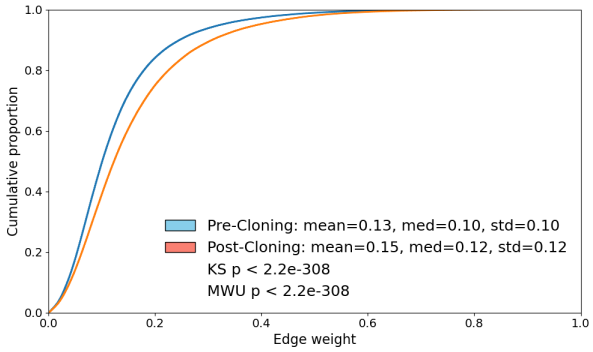


(b) WSU (BERT)

Figure 6.7: Comparison of pre- and post-cloning edge-weight distributions for WSU using LDA- and BERT-based cloning



(a) CSU (LDA)



(b) CSU (BERT)

Figure 6.8: Comparison of pre- and post-cloning edge-weight distributions for CSU using LDA- and BERT-based cloning

value. The descriptive statistics (mean, median, and standard deviation) and significance tests are reported within each plot to quantify the differences between the distributions. Across all institutions, both methods show statistically significant differences ($p < 0.05$) between pre- and post-cloned networks, with the post-cloned versions consistently exhibiting higher similarity values.

For the LDA-based approach, the rightward shift in the post-cloned distribution is modest, particularly for MSU and WSU, where the increase in mean edge weight is noticeable

but small. Nevertheless, the difference remains statistically significant, as the large number of edges yields extremely small p-values, even when the edge-weight distributions are closely aligned. By contrast, the BERT-based results exhibit a more pronounced rightward shift across all institutions. This improvement can be attributed to several factors. First, compared to LDA, BERTopic generated a larger number of clones per researcher, thereby amplifying the shift in the similarity distribution. Second, except for MSU, the overall topic count of the BERT-based models was higher than that of the LDA models, which influenced the shape and spread of the similarity distributions. In general, increasing the number of topics makes the topic space more fine-grained and can lead to lower similarity scores overall due to greater sparsity. This explains why MSU, where the total number of topics was relatively similar between the two models, shows comparable mean edge weights for both approaches.

Overall, the results indicate that cloning prolific researchers enhances the topical similarity structure of the network, regardless of the method used. However, the magnitude of improvement varies, with BERT-based cloning producing a stronger and more consistent shift due to its contextual embeddings and higher topical granularity. In the following section, we examine how these changes in similarity distributions affect the broader network topology and community structures.

Table 6.3 summarizes the community statistics for MSU, WSU, CSU, and the combined MWC network using both LDA- and BERT-based cloning. Communities were detected with the NH-Louvain algorithm using a minimum community size threshold of 30, as described in the experimental design section. Across all institutions, BERT produced a larger number of smaller communities compared to LDA. For example, at MSU the total number of communities increased from 27 under LDA to 44 under BERT, while the median size dropped from 10 to 6. Similar patterns appear at WSU and CSU, reflecting that BERT generated more topic groups and created additional clones for prolific researchers, which led to smaller

Table 6.3: Community statistics across institutions and models

Metric	MSU		WSU		CSU		MWC	
	LDA	BERT	LDA	BERT	LDA	BERT	LDA	BERT
# Communities	27	44	59	77	63	106	144	207
Max Size	26	57	24	26	28	28	28	30
Min Size	3	2	2	2	2	2	2	2
Mean Size	11.11	8.43	11.66	10.35	13.21	10.04	13.27	10.92
Median Size	10	6	11	9	13	9	12	10
Mean Density	0.86	0.78	0.96	0.86	0.96	0.90	0.97	0.88

but more numerous communities.

The maximum community sizes remain fairly consistent across institutions because of the minimum size constraint applied in NH-Louvain. Most of the largest communities approach the threshold of around 30 members. The main exception is MSU (BERT), where the maximum community size reaches 57, suggesting that this group was dense enough to be split further during the recursive steps of NH-Louvain. Minimum community sizes are also consistent across all networks, typically between 2 and 3, representing small research groups or pairs of closely related collaborators.

When examining the mean community densities, LDA-based networks show slightly higher average densities than BERT-based networks. This indicates that LDA communities are more tightly connected on average. In contrast, BERT produces more communities with lower internal density, which is expected given the larger number of topics and clones it generates. As a result, the overall network becomes more fine-grained, with smaller and more focused groups but weaker average connections. The combined MWC network follows the same pattern, showing a consistent difference between the two approaches in both the

Table 6.4: Summary of overlapping researcher memberships across institutions and models

Metric	MSU		WSU		CSU		MWC	
	LDA	BERT	LDA	BERT	LDA	BERT	LDA	BERT
# Multi-comm	19	41	55	83	62	122	174	240
Max overlaps	4	10	6	6	5	8	6	9
Min overlaps	2	2	2	2	2	2	2	2
Median overlaps	2	3	2	3	2	3	2	3

number and density of communities.

Table 6.4 summarizes overlapping researcher memberships across institutions where metric "# Multi-comm" specifies the number of researchers belonging to multiple communities. Similar to the community statistics, BERT-based networks consistently show a higher number of researchers appearing in multiple communities compared to LDA. For example, at MSU the number of overlapping researchers increases from 19 (LDA) to 41 (BERT), and at CSU from 62 to 122. This pattern is also evident in the combined MWC network, where overlapping memberships nearly double under BERT. The increase reflects the finer topical grouping and additional clones created under BERT, which allow individual researchers to appear across several topical clusters rather than being concentrated in a single one.

Although the maximum overlap counts vary slightly by institution, most researchers with multiple memberships belong to only two or three communities, indicating moderate cross-domain connections rather than widespread redundancy. It is also important to note that overlapping memberships are not the same as the set of prolific researchers selected for cloning. In some cases, cloned researchers were assigned to the same community during the refinement step, which merged their clones into a single node. This means that while clones were generated based on document-level clustering, strong topical similarity among those

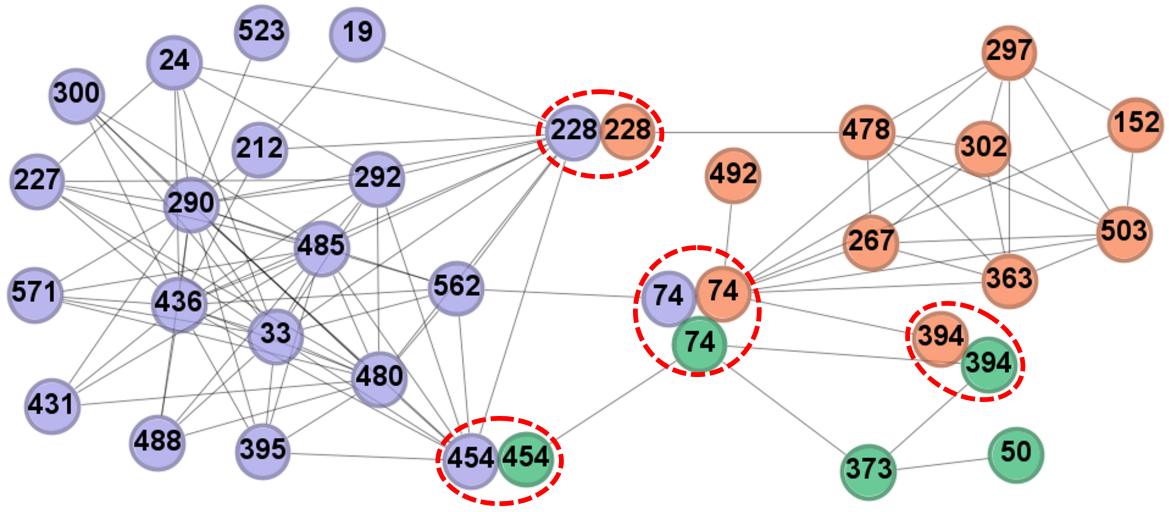


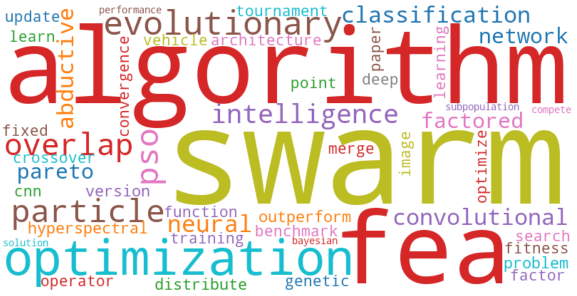
Figure 6.9: Overlapping communities in the MSU network (Clone-BERT). Dashed circles indicate overlapping memberships across multiple communities

clones kept them within the same community, resulting in fewer overlapping researchers than the total number of prolific ones cloned.

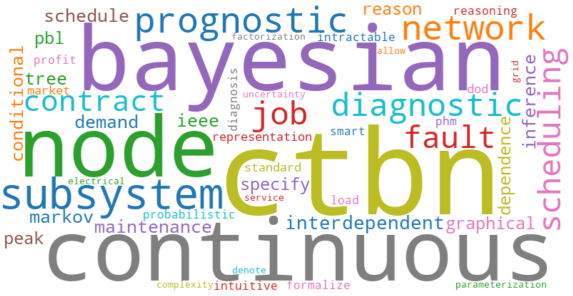
To further illustrate how cloning enables overlapping community membership, Figure 6.9 presents an example subnetwork from the MSU dataset following the Clone-BERT model. The visualization highlights researchers appearing in multiple communities, revealing the overlapping structure that emerges once prolific authors are represented through their cloned topical profiles. In the figure, dashed circles denote researchers whose cloned representations place them in more than one community, indicating overlapping membership across multiple topical areas.

As shown in Figure 6.9, four researchers participate in overlapping communities, with researcher 74 appearing in all three, while researchers 228, 394, and 454 belong to two each. Without cloning, traditional community detection would assign these individuals to a single dominant community based on their most frequent topical association. The cloning process, however, separates their publication clusters into distinct topical instances, allowing their multiple research directions to appear in different communities—capturing the

(a) WordCloud of Researcher 454 (pre-cloning)



(b) WordCloud of Clone 454–1



(c) WordCloud of Clone 454–2

Figure 6.10: Top 50 terms for Researcher 454 and its two clones under Clone-BERT

interdisciplinary nature of their work.

Figure 6.10 further demonstrates how cloning separates overlapping topical focuses. WordClouds are generated by weighting each term's probability by its topic probability and selecting the top 50 words for visualization. The pre-cloning WordCloud (Figure 6.10a) combines terms from both optimization and probabilistic modeling, reflecting a blended topical signature. After cloning, these subthemes are separated: Clone 454–1 emphasizes bio-inspired optimization and evolutionary computation, whereas Clone 454–2 highlights Bayesian modeling and temporal reasoning. This division allows the cloned researcher instances to appear in distinct communities, improving both interpretability and balance in the topic-based collaboration network.

6.5 Summary

In this chapter, we examined the publication imbalance problem in collaboration recommendation systems using scholarly community detection and social network analysis. Publication imbalance refers to the tendency of a small group of researchers to publish significantly more than others, which skews the overall publication distribution. Without properly addressing this imbalance, which has been largely overlooked in the literature, prolific researchers tend to remain confined within a single community representing their dominant research area. As a result, their secondary and less dominant topics become hidden. Since one of the goals of this dissertation is to promote interdisciplinary collaboration, neglecting these secondary themes limits potential recommendation opportunities that could foster cross-disciplinary connections.

To address this issue, we proposed a cloning-based strategy that creates multiple representations of prolific researchers to better capture the diversity of their research themes. Two cloning approaches were developed: an LDA-based method and a BERTopic-based method using a fine-tuned SciBERT model. These approaches generate multiple clones for each prolific researcher, allowing each clone to represent a distinct topical focus. Experimental results demonstrated that cloning improves the overall topical similarity structure in both approaches, particularly in the BERT-based model, and enhances community quality by distributing prolific researchers across multiple communities based on their thematic diversity. This, in turn, increases the likelihood of uncovering new potential collaboration links.

Several directions for future work remain. First, since the cloning approach naturally leads to overlapping community memberships, an important question is whether existing overlapping community detection algorithms could produce similar structures. Although our current framework relies on discrete community detection, comparing its output to

overlapping methods could reveal useful insights, even if such methods may not directly address the imbalance problem.

Second, as our methods are unsupervised, evaluating their effectiveness remains a challenge. While both LDA- and BERT-based cloning showed promising results quantitatively and qualitatively, as demonstrated by the word cloud visualizations, determining which performs better is not straightforward. Each model exhibits strengths in different areas: LDA tends to produce stronger similarity scores and tighter communities, whereas BERT captures a broader range of topics but with lower density and more clones. To further justify and interpret these differences, the next chapter focuses on evaluation and explainability, assessing the quality of the recommendations and providing justification for the underlying model behavior.

CHAPTER SEVEN

EVALUATION AND EXPLAINABILITY

In this chapter, we address the final research question of this dissertation: "How effective is the proposed collaboration recommendation framework, and how can we provide interpretable explanations for recommended connections in scholarly networks?" While the previous chapters focused on constructing topic-based scholarly networks, modeling hierarchical community structures, and addressing publication imbalance through cloning, these contributions must be evaluated to ensure that they can support reliable and explainable recommendations.

Scholarly collaboration evolves as researchers move into new or emerging areas. Methods based solely on observable connections, such as coauthorship networks, tend to struggle in these settings because researchers with limited or no prior collaborations appear as isolated nodes, a classic manifestation of the cold-start problem. In contrast, topic-based similarity constructs connections directly from publication content, allowing potential collaborators to be identified even when social links are missing. While this content-driven approach is less susceptible to the traditional cold-start issue, its effectiveness still depends on how well the system handles sparse or incomplete publication histories and whether the resulting similarities reflect genuine collaboration potential. In addition, users require clear, interpretable explanations for why a recommendation is made to support trust and adoption in real decision-making contexts.

This chapter evaluates the proposed framework from two complementary perspectives. First, we examine the stability of topical similarity by withholding portions of coauthored publications during model training. This simulates cases where researcher information is incomplete or where a new researcher enters the system with no prior coauthorship history.

Second, we assess recommendation accuracy using coauthorship as a form of ground truth validation to determine whether the system can retrieve known collaborators under held-out conditions. Finally, we explore explanation methods that highlight the topical evidence supporting each suggested collaboration.

This chapter begins with the motivation and problem statement, followed by related work. We then describe the evaluation methods for topical similarity and collaboration recommendation, and present experimental results. The chapter concludes with case studies and overall observations that demonstrate the practical utility of the proposed framework.

7.1 Motivation and Problem Statement

Identifying suitable collaboration partners is challenging, especially in large and interdisciplinary research environments. Coauthorship-based approaches often reinforce existing collaborations rather than reveal new ones. Topic-based similarity offers a more flexible approach by focusing on shared research interests.

However, the effectiveness of topic-based recommendations depends on two key factors. First, researchers may have limited or no coauthorship history, especially early in their careers. A reliable framework must still provide reasonable similarity estimates when relevant publications are missing, relying solely on the researcher's topical contributions. Second, similarity alone does not always provide confidence in the recommendation. Explanations that highlight shared topics can help users understand and trust why a suggested collaboration makes sense.

These considerations motivate the need to evaluate both the stability and interpretability of the proposed recommendation system. In this chapter, we investigate:

- **Stability:** How well are meaningful similarity relationships preserved when coauthored publications are withheld?

- **Accuracy:** How effectively can the system identify known collaborators under held-out evaluation?
- **Interpretability:** How can topical information be presented to show why a recommendation is made?

Problem Statement: Given a topic-based scholarly network enhanced by cloning to address data imbalance, how do we evaluate whether the framework can reliably recover topical similarity, accurately identify potential collaborators, and provide clear explanations for recommended collaborations?
Hypothesis: We hypothesize that the cloned representation will improve robustness to sparse or incomplete publication histories by capturing secondary topical themes and that visualization-based explanation will support interpretability for real users.

7.2 Related Work

Collaboration recommendation research has traditionally relied on coauthorship networks, where existing relationships are used to identify future collaboration potential [5, 81]. This approach supports accurate retrieval of established partnerships, but it tends to reinforce known structures rather than uncovering new or interdisciplinary opportunities. Because many impactful collaborations occur outside prior coauthorship history, structural approaches alone are not sufficient for exploring early-stage or emerging research relationships.

Content-based and hybrid methods have therefore been introduced to incorporate topical similarity when predicting collaboration links. In scholarly networks, link prediction refers to estimating whether a pair of researchers is likely to collaborate in the future based on

current network structure or content features. Topic modeling and semantic similarity have shown effectiveness in link prediction tasks by identifying researchers with related expertise even when coauthorship information is sparse [45, 95, 103, 132]. These methods help mitigate the cold start problem [109], but evaluation practices typically continue to rely on past collaborations as ground truth. Metrics such as precision and recall are widely used [66, 74], yet they inherently emphasize known and often intra-domain relationships while overlooking novel or cross-community recommendations that lack historical validation.

Recent research highlights the importance of recommendation diversity and interdisciplinarity in scholarly networks, arguing that evaluation should consider the breadth of opportunities surfaced rather than accuracy alone [80, 131]. These works demonstrate that systems capable of identifying relevant collaborators outside existing domains may better support scientific innovation and broader institutional goals.

Explainability has also become a critical design consideration in recommendation systems, particularly when recommendations involve new or nontraditional connections. Visualization-driven explanations, such as topical alignment or shared keyword evidence, have been shown to improve transparency and user trust in scientific literature recommendation [36]. Broader surveys on trustworthy and transparent recommenders further emphasize interpretability as an essential component for deployment in real-world decision making [15, 123, 124]. Despite these advances, many collaboration recommenders still operate as black boxes, offering limited interpretability when recommending new or interdisciplinary relationships.

In summary, prior work has demonstrated the strengths of topic-based collaboration prediction and trustworthy recommendation practices, yet evaluation strategies still heavily favor predictions based on past collaborations and provide limited insight into interdisciplinary potential. This chapter builds on existing approaches by examining the stability of topical similarity under reduced publication information, validating recommendations

using held-out collaboration links, and incorporating topical explanations to clarify why a researcher is recommended. These considerations support the intended use of our framework to uncover not only accurate but also interpretable and potentially interdisciplinary collaboration opportunities.

7.3 Stability of Topical Similarity

Understanding how well topic-based similarity holds when publication information is incomplete is necessary for reliable collaboration recommendations. To examine this, we simulate realistic information sparsity by removing shared co-authored publications and evaluate whether similarity relationships remain stable when explicit collaboration signals are missing.

7.3.1 Removing Shared Co-Authored Publications

When two researchers coauthor one or more publications, those shared documents may inflate topic-based similarity scores due to direct textual overlap. An important question is how much topic similarity changes when this overlapping content is removed. This is especially relevant in real-world scenarios where the goal is to recommend potential collaboration partners even if no prior coauthorship exists. One of the motivations behind our content-based approach is that early-career researchers with limited publication history should still be able to discover collaboration opportunities based solely on alignment of topical interests.

To evaluate this, we measure the stability of similarity scores by removing the publications coauthored by each researcher pair and regenerating a trimmed model, meaning a version of the topic model trained without the shared coauthored works for that pair. We then recompute the researcher–topic representations and JSD similarities and compare them against the full model that retains all publications. If the resulting similarity remains

comparable to the full model, it suggests that the relationship is supported by broader topical alignment rather than direct coauthorship alone. Conversely, substantial similarity drops would indicate that the relationship relies primarily on previously shared publications.

We apply this analysis to the MSU, WSU, and CSU datasets (see Chapter 3 for details). For each institution, we identify all coauthored publications and construct trimmed corpora, meaning versions of the text corpus where the coauthored works for each researcher pair have been removed. In addition, we use the combined MWC dataset to evaluate cross-institutional researcher pairs, allowing us to assess stability when collaborations span multiple universities. Although most academic works involve coauthorship, restricting the researcher population to current faculty allows us to remove shared publications without entirely eliminating the document history for most authors. We then apply both topic modeling strategies under study: LDA and BERT-based BERTopic.

For LDA, we retrain the topic model on each trimmed dataset so that the model no longer has access to the content from the removed publications (see Section 4.3.1 for training details). To ensure comparability with the full model, we fix the number of topics to match the selected topic count from full data experiments. Since researcher similarity is computed from topic-probability distributions, preserving dimensional alignment across runs is critical for consistency.

For BERTopic, retraining introduces additional challenges. BERTopic relies on a sentence-embedding model followed by HDBSCAN clustering, meaning that both cluster structure and topic dimensionality can shift when documents are removed (see Section 6.3.1). To maintain a stable embedding space and isolate the effect of missing publication content, we do not fine-tune the SciBERT embedder or re-estimate topic clusters with the trimmed documents. Instead, we retain the full BERTopic model and use `transform()` to infer topic distributions for the trimmed publication sets. This ensures that any differences in author–topic vectors arise solely from the removal of coauthored publications rather

Table 7.1: Stability of topic-based similarities under trimming for each institution

Model	Institution	n_{pairs}	Mean ΔSim	SD ΔSim	Cohen's d	p-value
LDA	MSU	573	0.0496	0.0821	0.60	< 0.001
	WSU	1118	0.0560	0.0854	0.65	< 0.001
	CSU	2042	0.0595	0.0817	0.73	< 0.001
BERT	MSU	573	0.0597	0.1006	0.59	< 0.001
	WSU	1118	0.0690	0.1143	0.60	< 0.001
	CSU	2042	0.0653	0.1054	0.62	< 0.001

than from changes to the underlying embedding model or topic geometry. This approach is consistent with BERTopic's formulation, where the sentence embedder is expected to generalize to unseen or modified text without retraining.

To quantify stability, we compute pairwise similarity differences between full and trimmed models for all coauthored researcher pairs. We report the mean and standard deviation of the similarity differences, along with Cohen's d effect size to assess practical significance. In addition, we perform significance testing using the Wilcoxon signed-rank test [127] to determine whether similarity distributions differ beyond random variation. Cohen's d complements the hypothesis testing by capturing how large the differences are relative to overall similarity variability.

7.3.2 Single-Institution Comparison

Table 7.1 summarizes the stability results for each institution using both LDA and BERT. The number of evaluated pairs n_{pairs} varies by institution based on the prevalence of coauthorship. The mean similarity differences range from approximately 0.05 to 0.07, indicating only minor changes in researcher similarity when coauthored publications are

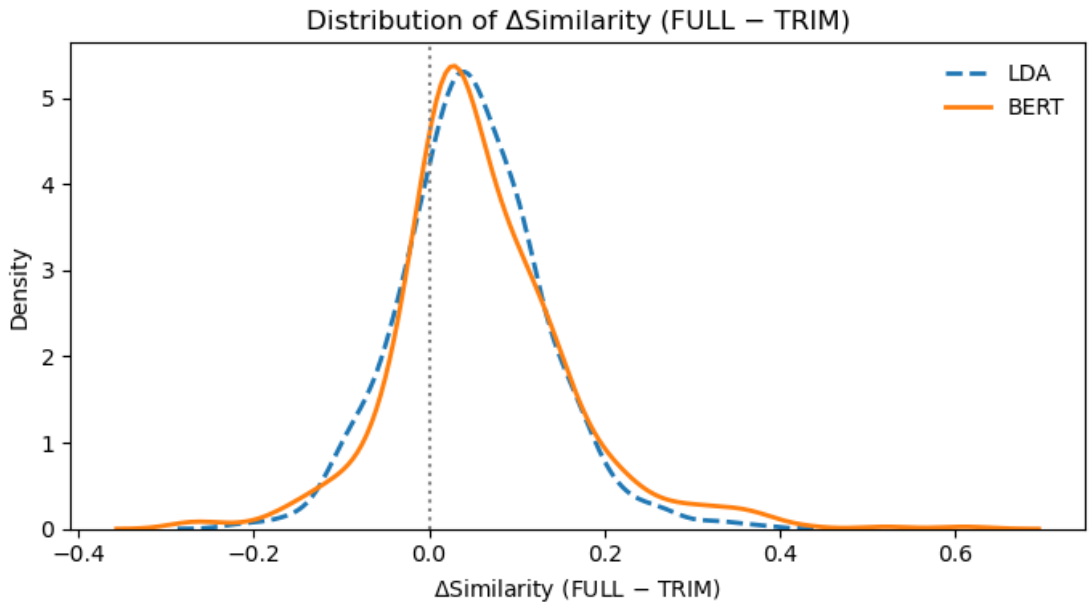


Figure 7.1: Distribution of similarity differences for the MSU dataset

removed. All comparisons show statistically significant differences (p-value < 0.001), which is not surprising given the large number of pairwise observations.

However, Cohen's d values fall in the small-to-moderate range (0.59–0.73), suggesting that the practical differences are limited relative to the total similarity scale. Under conventional interpretations of effect size, where $d \approx 0.2$ is considered small, 0.5 moderate, and 0.8 large [19], these results indicate that trimming affects similarity but not to a degree that substantially alters researcher relationships. BERT exhibits slightly stronger stability compared to LDA in some cases, particularly at CSU where the LDA effect size approaches the moderate–high threshold.

Figure 7.1 shows the distribution of similarity differences computed for MSU co-author pairs using LDA and BERT. Both curves are tightly centered close to zero, indicating that most similarity relationships remain largely unchanged when shared co-authored publications are removed. A slight positive shift is visible, meaning that similarity scores decrease

Table 7.2: Stability of topic-based similarities under trimming across institution pairs

Model	Inst Pair	n_{pairs}	Mean ΔSim	SD ΔSim	Cohen's d	p-value
	MSU-WSU	33	0.0211	0.0595	0.35	0.050
LDA	MSU-CSU	25	0.0250	0.0556	0.45	0.034
	WSU-CSU	79	0.0327	0.0564	0.58	< 0.001
	MSU-WSU	33	0.0258	0.0358	0.72	< 0.001
BERT	MSU-CSU	25	0.0134	0.0212	0.63	0.001
	WSU-CSU	79	0.0479	0.0751	0.64	< 0.001

somewhat when direct co-author evidence is removed, as expected. However, the narrow spread of both curves suggests that topic-based similarity is strongly supported by broader topical context rather than only the shared publications. Overall, the plot visually reinforces the small mean differences and modest effect sizes reported in Table 7.1, demonstrating stable behavior for both topic modeling approaches.

Overall, these findings indicate that within each institution, removing shared publications does not substantially distort topical similarity relationships. This supports the usefulness of topic-based collaboration signals for early-career researchers who might lack a large coauthorship record.

7.3.3 Cross-Institution Comparison

We next evaluate cross-institution co-authored pairs to examine whether topical similarity remains reliable for predicting collaborations beyond institutional boundaries. Cross-institution collaboration rates are considerably lower than within institutions, leading to much smaller sample sizes and therefore reduced statistical power. Nevertheless, the same trimming methodology is applied, and stability indicators are computed for MSU-WSU,

MSU-CSU, and WSU-CSU pairs.

Table 7.2 shows that mean similarity differences for cross-institution pairs remain small across both models (approximately 0.02–0.05), which again suggests strong topical stability. Cohen's d values are small-to-moderate (0.35–0.72), indicating that the practical effect of removing shared papers is minimal. Some tests do not achieve strong statistical significance due to the limited number of pairwise observations, but the direction and magnitude of changes remain consistent with the institutional results.

Taken together, these results show that topic-based similarity remains stable even when co-authored publications are removed. This means the similarity signal is supported by broader topical alignment rather than only direct collaboration history. The same general pattern holds for cross-institution relationships, although the number of available co-authored pairs is smaller in those settings. This behavior is important because it demonstrates that topic-based similarity can still identify meaningful collaboration opportunities beyond institutional boundaries and for researchers with limited co-author history.

7.4 Evaluation of Collaboration Recommendation

Evaluating recommendation quality in this dissertation presents several unique challenges. Because the proposed framework is unsupervised, there are no predefined "correct" collaborations and the quality of recommendations cannot be measured using standard supervised evaluation protocols. Prior studies on scholarly recommendation have addressed this by using coauthorship as a proxy ground truth: if two authors have coauthored in the past, a system that recommends one to the other is considered correct [22, 59, 129]. However, this approach is only a partial measure of success in our setting. Our broader goal is to promote interdisciplinary and novel collaborations, not only to recover existing ones. Using past coauthorship as a benchmark may therefore underestimate the real utility

of our approach, since a model that consistently suggests new, previously unconnected but topically compatible pairs would score poorly under this metric.

Nevertheless, evaluating against known collaborations remains valuable for validating model soundness: researchers who have published together tend to share strong topical similarity, so a model that ranks these authors highly can be regarded as learning meaningful scholarly representations. This section, therefore, reports both how the models recover known coauthor pairs and how they behave under controlled holdout settings that simulate missing collaboration data.

7.4.1 Ground Truth Setup

The publication data for the three institutions, MSU, WSU, and CSU, include both publication content and coauthorship metadata from the OpenAlex dataset. This allows us to construct topic-based scholarly networks using text content while separately retaining coauthorship information for evaluation.

The most direct approach would be to select a subset of coauthors, remove all shared works between them, retrain topic models, and then evaluate whether the system can rediscover them. However, this "complete removal" strategy poses a major problem: many researchers (especially those with few publications) would lose all of their works and thus vanish from the network. As a result, a fully random holdout would disproportionately exclude low-publication authors and bias the evaluation toward only the most prolific researchers.

Instead, we adopt a fifty–fifty split holdout. For each coauthor pair (a, b) with multiple shared publications, we randomly divide their joint works into two halves: half remain with author a, and half with author b. For example, if two researchers coauthored ten papers, five are retained for a and the remaining five for b, assigned at random. This partial holdout restricts direct coauthorship overlap while preserving both authors' topical representations.

It also avoids fully removing all publications for low-volume authors, which would eliminate their topical signal entirely and prevent meaningful similarity estimation. As a result, the evaluation can be conducted on the full author population rather than a small, filtered subset.

To reduce noise, we evaluate only pairs that share at least two coauthored works. Single-paper collaborations are excluded, as they may represent incidental or one-time collaborations rather than meaningful topical alignment.

We thus conduct two complementary evaluations:

1. **Full-data evaluation:** using all publications and coauthorships intact, to establish upper-bound performance; and

2. **Holdout evaluation:** using the 50/50 split, to test how well models recover hidden coauthors when shared publications are removed.

7.4.2 Models Evaluated

To evaluate the effectiveness of the proposed topic-based recommendation framework, we compare its performance against a strong lexical baseline. In total, five models are assessed, grouped into one baseline and four topic-based variants.

Baseline Model (TF–IDF). This baseline uses lexical similarity between researchers based on TF–IDF representations of their publications. Each document is encoded as a TF–IDF vector in the vocabulary space, and a researcher's representation is obtained by averaging and normalizing the TF–IDF vectors of their works. Researcher similarity is computed using cosine similarity. Although TF–IDF lacks the ability to capture latent or contextual semantics, it serves as a strong baseline for recovering existing collaborations because it directly reflects word-level overlap in research topics.

Proposed Topic-Based Models.

- **LDA.** This model is a probabilistic topic model in which each document is represented as a mixture of latent topics. Researcher-level representations are computed by aggregating the topic distributions of their publications, and topical similarity between researchers is derived from these distributions.

- **Clone–LDA.** This model extends the LDA framework by creating multiple cloned profiles for prolific researchers. Each clone receives a distinct subset of the researcher's publications, which enables the model to represent secondary or diverse topical interests that may be overshadowed in a single aggregated representation.

- **BERT.** This model refers to the BERTopic framework, which generates contextual document embeddings using transformer-based representations. Researcher vectors are obtained by averaging these embeddings across a researcher's publications, allowing the model to capture richer semantic relationships than TF–IDF or LDA.

- **Clone–BERT.** This model applies the same cloning strategy to the BERT-based workflow. Prolific researchers are partitioned into multiple cloned profiles, each representing a different subset of publications, which results in more balanced and interpretable researcher-level embeddings.

The topic-based models construct a topic-similarity graph among researchers, where edges represent pairwise topical similarity. Recommendations are generated using the Personalized PageRank (PPR) algorithm [90], which simulates a random walk over the researcher similarity network. Starting from a given researcher, the algorithm repeatedly moves to neighboring nodes based on edge similarity weights while occasionally restarting from the original node with probability α. This process assigns higher scores to researchers that are more frequently visited during the walk, effectively ranking those who are both topically related and structurally close within the network. We tested several α values and

fixed $\alpha = 0.1$, which provided stable results across institutions. The advantage of PPR is that it can surface indirect but topically coherent collaborators who are not directly connected, reflecting the potential for new interdisciplinary ties.

In contrast, the TF–IDF baseline does not use a graph structure. Recommendations are obtained by directly sorting cosine similarities between researchers. TF–IDF based approaches have been widely used in earlier scholarly recommendation systems (e.g., [111, 133]) and serve as a strong benchmark for recovering existing collaborations. However, they are limited in discovering latent or cross-domain connections and provide little interpretability beyond lexical overlap.

7.4.3 Evaluation Metrics

We adopt two standard metrics from the recommender-systems literature: **Hits@k** and **Mean Reciprocal Rank (MRR)**.

Hits@k [40] measures the proportion of queries (authors) for which at least one of the ground-truth collaborators appears among the top k recommendations:

$$\text{Hits@k} = \frac{1}{|Q|} \sum_{q \in Q} \frac{|A_q \cap R_q^{(k)}|}{|A_q|}$$

where Q is the set of query authors, A_q is the set of actual coauthors of author q, and $R_q^{(k)}$ represents the top-k recommendations generated for q. The metric measures the fraction of an author's true collaborators that appear among the top-k recommendations, averaged over all evaluated authors. In other words, Hits@k captures how effectively the model retrieves known collaborators within its top-k suggestions.

For example, if an author's top-5 recommendations include one of their real coauthors, that author contributes 1.0 to Hits@5; if not, 0.0. The metric thus reflects "recall", how often known collaborators appear within the top portion of the ranked list. We report Hits@3,

@5, @10, and @20 to examine how retrieval quality changes with list length.

MRR [40] complements Hits@k by assessing "how high" the true collaborator appears in the ranking. For each author q, let r_q denote the rank position of the first ground-truth collaborator in the full recommendation list. The reciprocal rank is $1/r_q$ if found, and 0 if none are retrieved. Then,

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{r_q}.$$

For example, if the first true collaborator appears at rank 2, the reciprocal rank is 0.5; if at rank 5, it is 0.2. MRR values therefore emphasize "early precision" where higher values indicate that correct collaborators are ranked near the top.

Together, Hits@k and MRR capture both the presence and the position of true collaborators in the ranked list, providing complementary views of model performance.

7.4.4 Results on Full Dataset

The evaluation was performed at the author level, considering only those researchers who had at least one coauthor with two or more shared publications. This filtering ensured that collaboration histories reflected meaningful partnerships rather than one-off papers. After filtering, 185 authors from MSU, 387 from WSU, and 519 from CSU were included, resulting in a combined evaluation pool of 1,633 researchers across all institutions.

Table 7.3 reports the results for each model across the three institutions and the combined dataset using the full publication data. The TF–IDF model achieved the highest overall scores in both MRR and Hits@k across all settings, reflecting its strength in capturing lexical overlap between coauthors. This outcome is expected because coauthored publications often share substantial vocabulary, phrasing, and domain-specific terminology, so a purely lexical model has a natural advantage when all shared publications remain in the dataset. TF–IDF can therefore recover known collaborators very effectively. However, its purely lexical nature limits its ability to surface new or cross-disciplinary connections, which are

Table 7.3: Performance of recommendation models on full datasets across institutions

Institution	Model	MRR	Hits@3	Hits@5	Hits@10	Hits@20
MSU	TF-IDF	0.59	0.68	0.79	0.92	0.96
	LDA	0.39	0.45	0.57	0.73	0.85
	Clone-LDA	0.40	0.47	0.59	0.76	0.89
	BERT	0.09	0.09	0.11	0.14	0.22
	Clone-BERT	0.20	0.23	0.29	0.35	0.47
WSU	TF-IDF	0.54	0.64	0.74	0.85	0.93
	LDA	0.37	0.43	0.54	0.68	0.80
	Clone-LDA	0.41	0.45	0.57	0.70	0.83
	BERT	0.30	0.33	0.44	0.57	0.68
	Clone-BERT	0.34	0.35	0.46	0.58	0.69
CSU	TF-IDF	0.51	0.56	0.67	0.81	0.90
	LDA	0.40	0.42	0.51	0.63	0.77
	Clone-LDA	0.41	0.43	0.52	0.66	0.79
	BERT	0.30	0.32	0.40	0.52	0.61
	Clone-BERT	0.30	0.31	0.39	0.52	0.63
Combined	TF-IDF	0.46	0.53	0.63	0.75	0.85
	LDA	0.36	0.39	0.46	0.59	0.71
	Clone-LDA	0.38	0.40	0.48	0.62	0.74
	BERT	0.23	0.25	0.32	0.41	0.50
	Clone-BERT	0.21	0.25	0.31	0.40	0.49

central to this dissertation's objectives.

LDA-based models demonstrated strong and consistent performance across all institu-

tions. The probabilistic topic modeling framework captures underlying thematic structures that extend beyond direct word matching, leading to meaningful collaboration retrieval even in cases where terminology differs. The Clone–LDA variant further improved results, suggesting that the cloning strategy successfully mitigates the publication imbalance problem by allowing prolific researchers' secondary topical areas to be more accurately represented. The benefit of this approach is particularly evident at MSU and the combined dataset, where Clone–LDA closely approaches TF–IDF in overall performance.

In contrast, BERT-based models underperformed in this evaluation. Both the standard and cloned BERTopic networks yielded lower MRR and Hits@k values, which aligns with the expectation that contextual embeddings capture broader semantic similarity rather than direct lexical overlap. Consequently, they are less effective for recovering past coauthor relationships but are expected to excel in identifying novel or interdisciplinary recommendations, which is explored later in this chapter.

Across institutions, the relative ranking of models remains consistent, though the performance margins vary. WSU and CSU exhibit narrower differences between TF–IDF and LDA-based approaches, suggesting smaller or more thematically cohesive author networks, while MSU shows the most pronounced gains from the cloning strategy. When all institutions are combined, TF–IDF remains the strongest performer, yet both LDA and Clone–LDA maintain robust ranking stability, demonstrating that topic-based representations generalize well across institutional boundaries.

7.4.5 Result on Holdout Dataset

To further examine the robustness of each model, we evaluated their performance under the 50/50 holdout setup, where half of the coauthored publications for each researcher pair were randomly removed. This setup intentionally weakens lexical similarity between coauthors, providing a stronger test of each model's ability to infer collaboration potential

Table 7.4: Performance of recommendation models on holdout datasets across institutions

Institution	Model	MRR	Hits@3	Hits@5	Hits@10	Hits@20
MSU	TF-IDF	0.39	0.45	0.57	0.72	0.83
	LDA	0.31	0.35	0.45	0.61	0.76
	Clone-LDA	0.31	0.37	0.48	0.63	0.78
	BERT	0.09	0.07	0.08	0.13	0.18
	Clone-BERT	0.19	0.20	0.25	0.32	0.42
WSU	TF-IDF	0.40	0.46	0.57	0.70	0.81
	LDA	0.28	0.32	0.43	0.56	0.70
	Clone-LDA	0.32	0.36	0.45	0.60	0.74
	BERT	0.21	0.22	0.31	0.45	0.56
	Clone-BERT	0.26	0.27	0.34	0.46	0.59
CSU	TF-IDF	0.39	0.42	0.52	0.66	0.77
	LDA	0.31	0.33	0.42	0.55	0.69
	Clone-LDA	0.32	0.34	0.44	0.56	0.70
	BERT	0.26	0.27	0.34	0.44	0.54
	Clone-BERT	0.26	0.27	0.33	0.44	0.56
Combined	TF-IDF	0.21	0.22	0.28	0.36	0.44
	LDA	0.26	0.27	0.35	0.46	0.59
	Clone-LDA	0.27	0.28	0.36	0.47	0.61
	BERT	0.17	0.18	0.24	0.31	0.40
	Clone-BERT	0.16	0.18	0.22	0.30	0.39

from limited information. Table 7.4 presents the results across all institutions and for the combined dataset.

As expected, overall performance values decrease compared to the full-data setting due to the reduced shared information between coauthors. The TF–IDF baseline experiences the largest drop in MRR and Hits@k, highlighting its heavy reliance on direct word overlap between publication texts. When coauthored papers are partially removed, TF–IDF struggles to recognize remaining connections because its representation is purely lexical.

In contrast, topic-based models show greater resilience to the holdout condition. Both LDA and Clone–LDA maintain substantially higher MRR and Hits@k values than their embedding-based counterparts, demonstrating that topic distributions capture latent thematic structures that persist even when shared papers are withheld. The cloning strategy again improves performance, especially in MSU and WSU, confirming that the diversified representation of prolific researchers helps preserve their topical breadth when data is reduced.

The most notable result arises in the combined dataset, where Clone–LDA surpasses TF–IDF across all evaluation metrics, achieving an MRR of 0.27 compared to TF–IDF's 0.21. This marks the first case where a topic-based network outperforms the lexical baseline, suggesting that topic-level abstraction and cloning together generalize better across diverse institutional domains. In this setting, TF–IDF's dependence on shared vocabulary becomes a limitation, whereas LDA models leverage cross-institutional topical overlap to recover meaningful collaborative links that are not driven by identical wording.

Figure 7.2 compares performance on the combined network under full and holdout settings using MRR and Hits@10. TF–IDF shows the largest decline after removing shared publications, reflecting its reliance on lexical overlap. LDA and Clone–LDA are more stable, consistent with topic distributions retaining latent thematic similarity when shared papers are reduced. The cloned variant exhibits the smallest degradation across both metrics, indicating that balancing publication volume preserves secondary topical signals. Under holdout, Clone–LDA surpasses TF–IDF, suggesting that topic-level abstraction and cloning

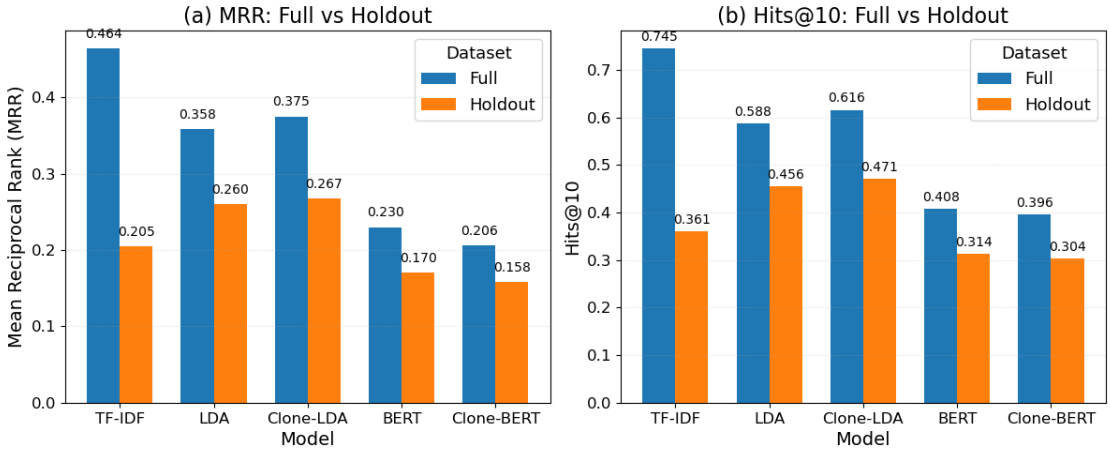


Figure 7.2: Full versus holdout performance on the combined dataset

generalize better across institutions when explicit word overlap is limited.

Overall, these results support the adaptibility of topic-based scholarly networks under information-sparse conditions. While TF–IDF remains competitive in settings with abundant shared data, LDA and particularly Clone–LDA demonstrate stronger generalization and stability when shared publications are restricted, reinforcing their suitability for discovering potential collaborations in partially observed or cross-domain environments.

7.5 Explainability of Recommendation

While topic models such as LDA and BERTopic provide interpretable topic–word distributions, the resulting collaboration recommendations are less transparent. Researchers may see a suggested collaborator but not understand why the connection was made, especially across institutions or disciplines. In practice, explainability is important not only for user trust but also for verifying that recommendations are driven by meaningful thematic overlap rather than superficial textual similarity.

TF–IDF–based models perform well in accuracy but offer limited interpretability, since they depend mainly on word co-occurrence rather than coherent topical structure. In

contrast, topic models can link recommendations to identifiable research themes, allowing users to trace how shared topics contribute to a suggestion. Even transformer-based models like BERTopic maintain this property to some extent, since topic–word distributions can still be extracted for interpretation. However, these distributions are derived indirectly from high-dimensional document embeddings rather than from an explicit generative process, which makes the semantic justification less transparent than in LDA. In BERTopic, the topic words are produced through the c-TFIDF transformation on embedding clusters, so they reflect patterns in the embedding space rather than clear probabilistic associations between words and topics.

Although every topic-model–based recommender proposed in this study can generate similar explanations through their underlying topic–word distributions, we focus the analysis on Clone–LDA as a representative case. Clone–LDA retains the interpretability of standard LDA while addressing publication imbalance through researcher cloning, allowing both dominant and secondary research themes to contribute to recommendation decisions. This makes the explanations more complete, as the model does not rely solely on a researcher's most frequent topics. In addition, Clone–LDA performed most consistently in the holdout evaluation, indicating that it captures more generalizable topical relationships rather than coauthorship-specific effects. For these reasons, we use Clone–LDA to illustrate how topic-based recommenders can produce interpretable explanations for why specific researcher pairs are suggested. The following sections describe the process used to generate these explainability visualizations.

7.5.1 Topic Distribution Comparison

To generate explainable visualizations, we first construct topic-based explanations for recommended researcher pairs using their underlying topic probability distributions. For each recommended pair, we take both researchers and use the trained Clone–LDA model

to obtain their topic probability distributions. Because this version of LDA includes cloned representations of prolific researchers, each clone captures a distinct subset of publications grouped by topical similarity. We compute the topic probability distribution for each clone by concatenating all documents associated with that clone into a single large document. This helps prevent secondary themes from being overshadowed by dominant research areas, making the resulting distribution more representative of the researcher's topical diversity.

When generating explanations, we handle cloned researchers by comparing every possible clone combination between the two researchers. If researcher A has three clones and researcher B has two, we compute the similarity score for all six possible pairs and select the pair with the highest topic similarity. This approach mirrors the actual recommendation process, where clone-level edges are weighted by their topical similarity within the network. Using the most similar clone pair therefore provides a consistent and justified basis for explanation.

After selecting the best-matching clone pair, we compute topic similarity across all topics by multiplying the two researchers' topic probabilities. The top five topics with the highest joint probability scores are then identified as the shared topics driving the recommendation. These topics are later used to generate shared wordclouds and a bidirectional bar chart that visualizes the probabilistic contribution of each researcher to the shared themes (for example, Topic 71 may have a probability of 0.30 for Researcher A and 0.20 for Researcher B).

we compute a topic similarity score based on the product of their topic probabilities and identify the top five shared topics with the highest joint probability. These topics represent the most influential topical areas driving the recommendation. The corresponding top words from each shared topic are then aggregated to form a shared wordcloud, illustrating the overlapping research themes that connect the two researchers.

7.5.2 Wordcloud Generation

The next phase of the explainability plots focuses on visualizing the topical themes that define each researcher and their shared research areas. Wordclouds provide an intuitive way to represent the relative importance of words within a topic or distribution. Each word's font size is proportional to its associated probability value, offering a quick visual impression of what the topic, or in this case the researcher's thematic profile, represents. Although wordclouds are often used to describe the content of individual topics discovered by LDA, here we adapt the same concept to illustrate both individual and shared topical structures between researcher pairs.

At the individual level, we first use each researcher's topic probability distribution and select their top five topics based on probability scores. The topic–word distributions are obtained from the trained LDA model, which assigns a probability to every word within a topic (for example, $machine$: 0.15, $algorithm$: 0.12, etc.). To derive a researcher-specific word distribution, we multiply the researcher's topic probability by the word probabilities of their top ten words in each of the top five topics, resulting in a weighted set of fifty words that best represent that researcher's overall topical profile.

We follow a similar approach for generating the shared wordclouds. Instead of using individual topic probabilities, we rely on the joint topic probabilities obtained from the multiplication of the two researchers' topic distributions. The top five topics with the highest joint probabilities are used to weight their corresponding topic–word distributions, producing a combined set of words that visually highlights the shared themes driving the collaboration recommendation.

7.5.3 Case Studies

To illustrate how the explainability framework supports the interpretation of collaboration recommendations, we present a set of case studies drawn from the MSU network. These

examples demonstrate how topic–distribution comparisons and wordcloud visualizations jointly reveal the topical rationale behind a recommendation. For each case, we include both individual and shared wordclouds, along with the bidirectional topic–probability plots described earlier, allowing direct visual comparison of topical alignment between researchers.

Our first set of case studies focuses on a single MSU researcher (Researcher 454) and examines three of their recommended collaborators. These examples include a top-ranked recommendation (rank 1), a mid-ranked recommendation (around rank 10), and an additional lower-ranked example. Together, they illustrate how shared topical structure becomes less pronounced as the recommendation rank decreases.

To complement these examples, we include a second set of case studies centered on a different MSU researcher. For this researcher, we present two recommendation examples, again drawn from the top-ranked and mid-ranked portions of their recommendation list. These cases demonstrate how the framework explains recommendations for researchers with different topical profiles and highlight the consistency of the explainability approach across individuals.

Taken together, these five examples provide a representative view of how the explainability framework reveals the key topics, shared themes, and topical alignment patterns that drive collaboration recommendations across diverse ranking positions and researcher profiles.

Figure 7.3 shows the top-ranked recommendation for Researcher 454, where the model identifies a strong alignment between the researchers' topical interests. The bidirectional bar chart reports a JSD topic similarity score of 0.63 and highlights several shared topics with high joint probabilities, particularly Topic 71, which exhibits the greatest shared weight. The shared wordcloud reinforces this pattern, displaying dominant terms such as "network", "feature", "classification", "machine", and "learning" (separated as unigrams during tokenization). Together, these terms indicate substantial overlap in machine learning–related research themes and suggest that both researchers engage with similar

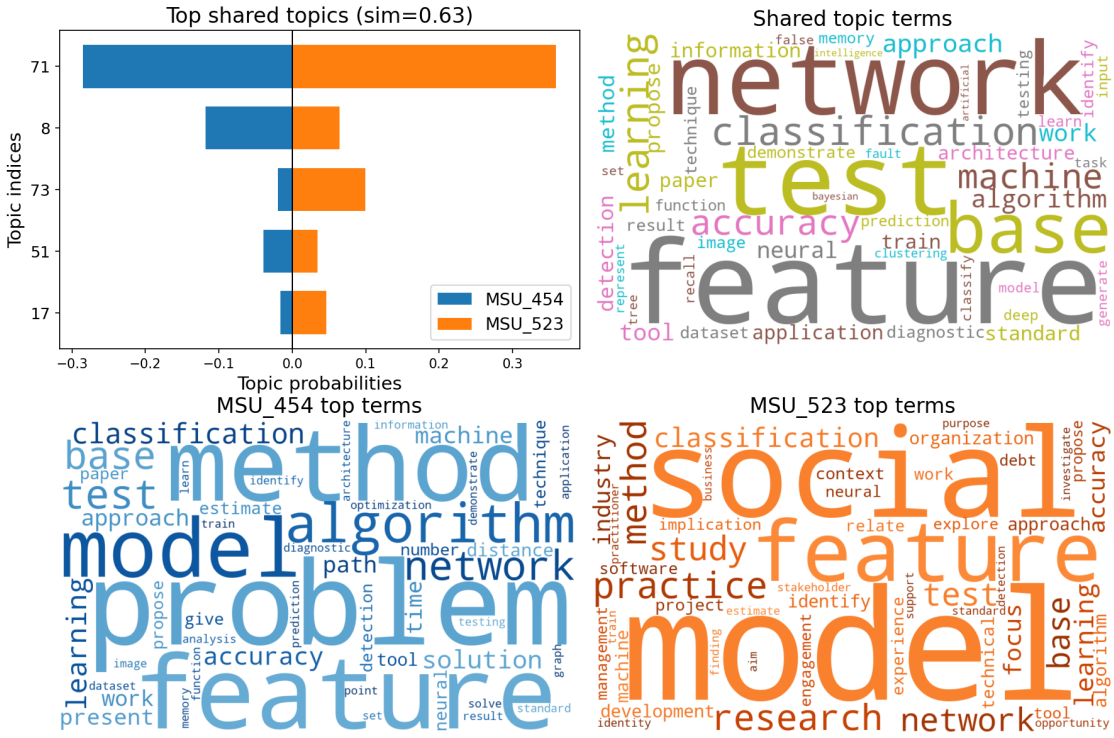


Figure 7.3: Explainability plot for the top-ranked recommendation of Researcher 454

modeling and predictive approaches.

At the individual level, both researchers share a focus on data-driven modeling and classification, but their topical emphasis differs. Researcher 454's wordcloud highlights terms such as "method", "problem", "feature", and "algorithm", reflecting a strong orientation toward model development and computational methodology. In contrast, the recommended collaborator emphasizes words such as "model", "social", and "practice", indicating a broader applied perspective where modeling is used to address contextual or behavioral problems. The overlap in shared topics such as "network", "test", and "classification" explains the high similarity score, while the differences in individual word distributions reveal complementary methodological and applied strengths that justify this top-ranked recommendation.

Figure 7.4 shows a mid-ranked recommendation for Researcher 454, where the overall topic similarity is moderate (0.47). Compared to the top-ranked pair, the overlap is narrower

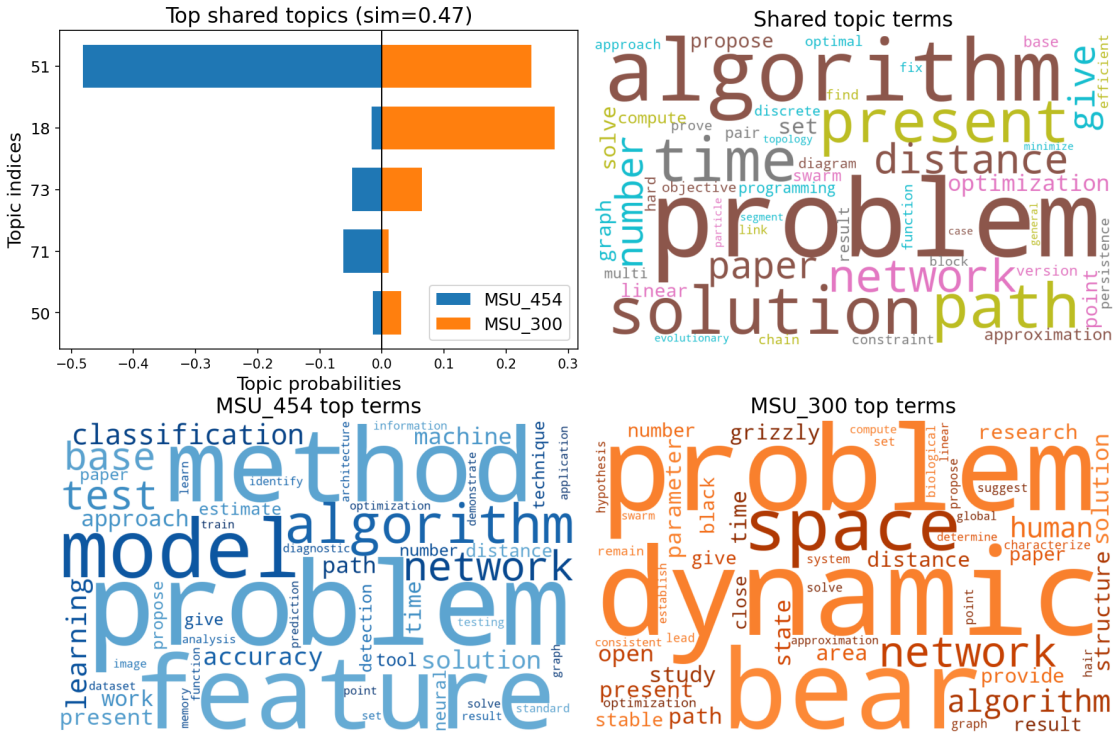


Figure 7.4: Explainability plot for a mid-ranked recommendation of Researcher 454

and focuses on more specific technical themes. The bidirectional bar chart highlights Topic 51 as the strongest shared topic, followed by several lower-weight topics with smaller joint probabilities. This indicates that the two researchers share common ground in computational and algorithmic methods, but their connection is limited to a few focused areas rather than broad topical coverage.

The shared wordcloud supports this pattern, showing dominant terms such as “algorithm”, “problem”, “solution”, “path”, and “time”. These words suggest an emphasis on optimization and graph-based reasoning, which differs from the broader machine learning focus seen in the top-ranked case. The collaborator’s wordcloud shows words like “dynamic”, “space”, “parameter”, and “network”, pointing to interests in mathematical modeling and dynamic systems. Together, these visualizations show that the model identified a link based on shared computational techniques, leading to a moderate yet meaningful similarity between

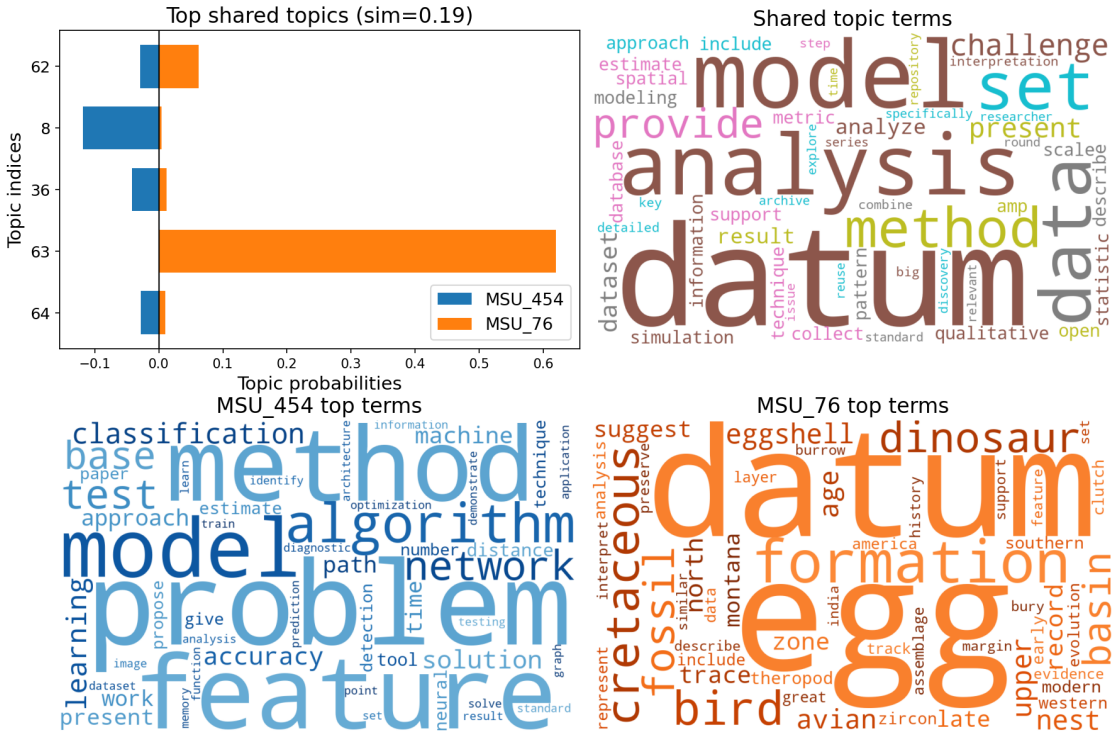


Figure 7.5: Explainability plot for a low-ranked recommendation of Researcher 454

the two researchers.

Figure 7.5 shows a low-ranked recommendation for Researcher 454 with an overall similarity score of 0.19, indicating weak topical alignment. The bidirectional bar chart clearly reflects this imbalance, as Topic 63 dominates the shared space while the remaining topics contribute very little. This pattern suggests that the recommendation is driven by a single broad topic rather than consistent overlap across multiple areas.

The shared wordcloud also shows limited overlap. Although both researchers use generic terms such as “data”, “model”, and “analysis”, these words appear in different research contexts. The recommended collaborator’s wordcloud highlights domain-specific terms such as “dinosaur”, “fossil”, “formation”, and “cretaceous”, pointing to a focus in geological sciences. In contrast, Researcher 454’s vocabulary remains centered on computational methods and modeling. The lack of overlapping specialized terminology and the strong

dominance of one researcher's topic profile explain the low similarity score and weaker relevance of this recommendation.

Having demonstrated how the explainability framework characterizes high-, mid-, and low-ranked recommendations for Researcher 454, we next examine a second focal researcher from the MSU network. This allows us to assess whether the interpretability patterns observed earlier generalize to researchers working in different topical domains. In this case, we focus on Researcher 74, who specializes in optical and remote sensing research.

Figure 7.6 shows the top-ranked recommendation for Researcher 74, with a high similarity score of 0.76. The bidirectional bar chart highlights several shared topics with substantial joint weight, particularly those centered on optical sensing and imaging. The shared wordcloud reinforces this pattern, with dominant terms such as "optical", "image", "polarization", "infrared", and "sky", indicating that both researchers work extensively with remote sensing instrumentation and environmental imaging.

Although the shared topical structure is strong, the individual wordclouds reveal subtle differences in emphasis. Researcher 74 places greater focus on measurement and systems-level instrumentation, with terms such as "laser", "lidar", and "sensor". The recommended collaborator displays similar domain themes but places more emphasis on application-oriented terms such as "temperature", "imaging", and "spectral". These complementary emphases fall naturally within the same research area and explain why the model ranks this collaborator near the top of the recommendation list.

Figure 7.7 presents a mid-ranked recommendation for Researcher 74, with a similarity score of 0.46. Compared to the top-ranked case, the overlap is narrower and reflects alignment primarily through broader methodological themes rather than tightly shared domain topics. The bidirectional bar chart shows that Topic 58 carries the greatest joint contribution, while the remaining shared topics have more modest weights, indicating that the connection is supported by a smaller set of common research directions.

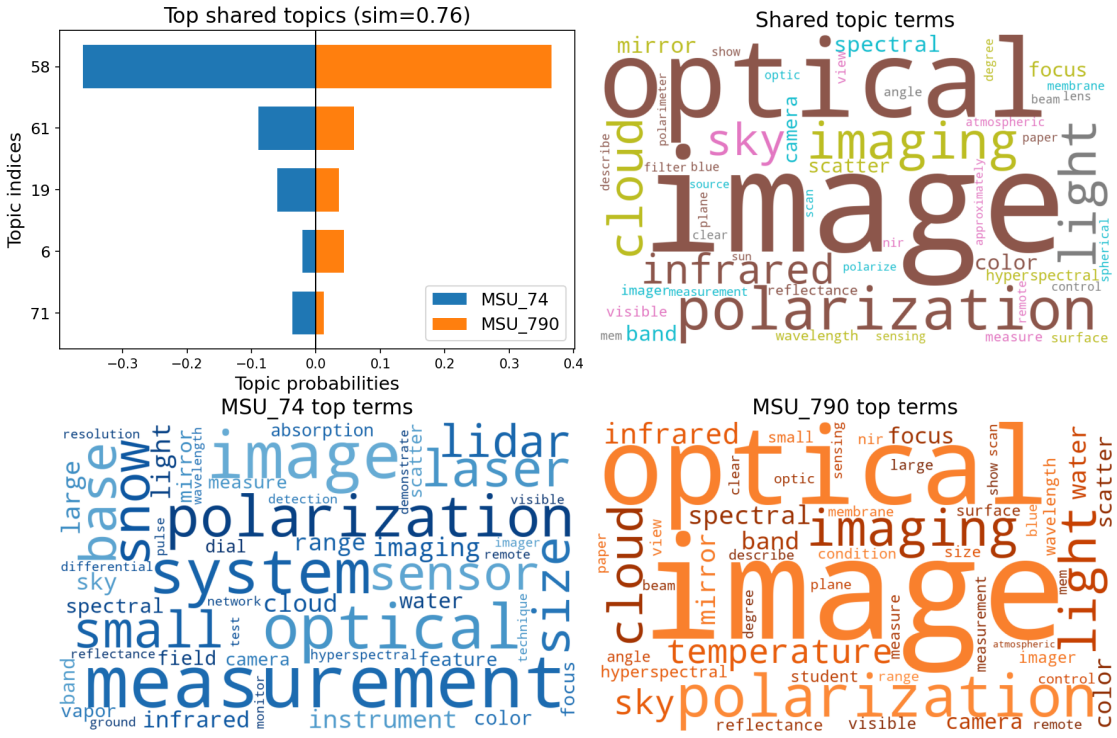


Figure 7.6: Explainability plot for a top-ranked recommendation of Researcher 74

The shared wordcloud reflects this partial alignment. While both researchers use terms such as “network”, “feature”, and “classification”, these appear in different topical contexts. Researcher 74’s work remains centered on optical and sensing themes, whereas the recommended collaborator’s vocabulary highlights machine learning applications, including terms such as “method”, “accuracy”, “learning”, and “neural”. Several of these terms are consistent with ML-based analysis of remote sensing data, which provides a methodological link between the two researchers. However, the collaborator’s emphasis on computational modeling differs from Researcher 74’s instrumentation-focused work, resulting in moderate topical alignment and a mid-ranked recommendation.

Across these examples, the explainability plots illustrate how topic-based similarity influences collaboration recommendations in Clone–LDA. High-ranked pairs show strong overlap across multiple topics and shared vocabulary, whereas mid- and low-ranked pairs

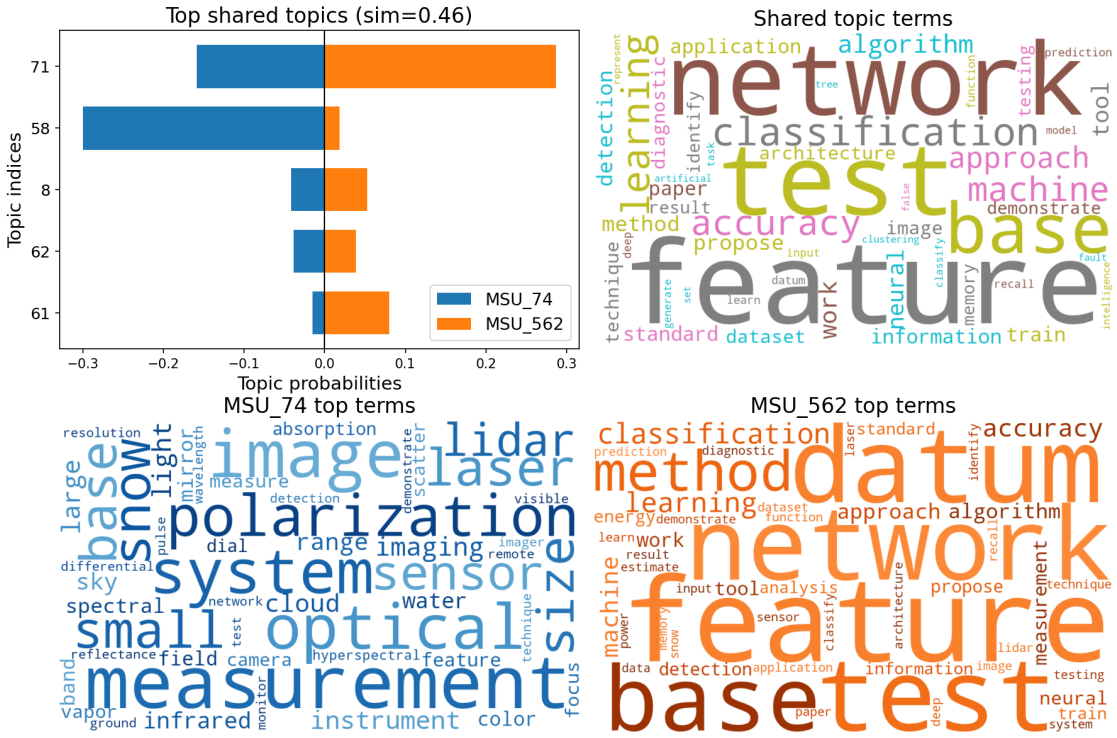


Figure 7.7: Explainability plot for a mid-ranked recommendation of Researcher 74

reveal gradually weaker topical alignment. The cases also demonstrate that this behavior is consistent across researchers working in different topical domains. These visual explanations improve transparency in the recommendation process, helping users verify that suggested collaborations arise from genuine topical relationships rather than surface-level textual overlap.

7.6 Summary

In this chapter, we evaluated the performance and reliability of our topic-based recommender models using historic coauthorship data. Two complementary perspectives guided this analysis: stability and accuracy. The stability experiments tested how robust topic-based similarity remains when portions of coauthored publications are withheld,

simulating realistic scenarios where researcher information may be incomplete or where new faculty members enter the network with limited publication history. The accuracy evaluation then measured how well the models could reproduce known coauthorships, using them as a form of ground-truth evidence for successful recommendations.

We compared both LDA- and BERT-based models along with their cloned counterparts to examine the effect of cloning on recommendation quality. The results showed that topic-based representations, particularly LDA and Clone–LDA, maintain strong performance even when publication overlap is removed, indicating that they capture generalizable thematic alignment rather than relying solely on prior collaborations. In contrast, lexical models such as TF–IDF showed greater sensitivity to missing documents, confirming that topical abstraction provides a more stable foundation for collaboration discovery.

Beyond predictive accuracy, this chapter also explored explainability as an essential aspect of model transparency. Using Clone–LDA as a representative example, we demonstrated how topic distributions and shared wordclouds can help visualize the underlying thematic evidence that drives recommendations. These explanations not only justify individual recommendations but also offer insights into how different topical strengths or complementary expertise contribute to meaningful collaboration suggestions.

Future work should extend these evaluation strategies by incorporating diversity-aware metrics that capture interdisciplinary potential rather than relying solely on historical coauthorship. Integrating such measures would provide a more comprehensive view of recommendation quality and better reflect the broader goal of promoting cross-domain and interdisciplinary research connections.

CHAPTER EIGHT

SCHOLARNODE APPLICATION

The previous chapters established a comprehensive framework for topic-based scholarly collaboration recommendation, encompassing topic modeling, hierarchical community detection, imbalance handling, and evaluation of model stability and explainability. While these experiments validated the framework empirically, they remain primarily analytical. To bridge the gap between theoretical validation and practical usability, we developed an operational prototype system titled "ScholarNode." The goal of ScholarNode is to transform the underlying research framework into an interactive web platform that enables exploration of scholarly networks, visualization of topical communities, and generation of collaboration recommendations in an interpretable manner.

ScholarNode was designed around three core principles: interpretability, modularity, and transparency. Rather than relying solely on co-authorship frequency or citation metrics, it centers on topic-driven similarity derived from probabilistic and transformer-based models. This approach enables the identification of latent interdisciplinary connections that might not be visible through traditional collaboration networks. The system thus serves as both a demonstration and validation of the dissertation framework's practical applicability in real-world academic ecosystems.

By providing an accessible interface to navigate researcher relationships and topical structures, ScholarNode facilitates data-driven decision-making for research administrators, interdisciplinary initiatives, and individual scholars seeking potential collaborators. The platform illustrates how topic-based community modeling and similarity analysis can be operationalized into a functional tool for exploratory analysis of scholarly activity.

8.1 System Overview

ScholarNode integrates the core conceptual components developed across RQ1–RQ4 into a unified and interactive framework. It builds on the topic modeling and network construction pipelines (Chapter 4), hierarchical community detection mechanisms (Chapter 5), and the data imbalance strategies introduced in Chapter 6, along with the evaluation and explainability measures from Chapter 7. Although the current prototype implements the standard LDA and BERT-based topic networks, the system architecture is designed to accommodate cloned topic models in future iterations. The prototype was developed for Montana State University but can be readily extended to other institutional contexts.

Internally, the system connects several sequential processes: topic distributions from LDA and BERT models are used to compute document-level and researcher-level similarities, which are then aggregated to construct topic-based networks. These networks are passed through hierarchical community detection algorithms, producing multi-scale community structures. The resulting network indices are stored in a database and exposed through a lightweight API that supports the web interface. Recommendations are generated by applying the Personalized PageRank algorithm, seeded on the topical distribution of a selected researcher to rank potential collaborators within and beyond the immediate community.

At the user level, ScholarNode offers three primary functionalities: a search interface, an individual researcher profile view, and a topical concept browser. Users can query by researcher name or by domain-specific keywords such as “agriculture,” “machine learning,” or “forest fire,” and are directed to relevant researcher profiles. Each profile page presents a force-directed graph visualization, which arranges vertices by simulating attractive and repulsive forces so that closely related researchers appear near each other while unrelated vertices are pushed farther apart. Supplementary panels display individual and community-

level wordclouds summarizing dominant research themes, publication histories, and a list of both local and external collaborators with clickable links to their profiles. Recommendation scores are shown alongside collaborator suggestions, reflecting computed similarities and network-based influence.

Although the current prototype focuses on LDA and BERT-based topic networks, it operationalizes all essential components needed to support search, community exploration, and topic-based collaboration recommendations. The modular structure of the system allows additional capabilities to be incorporated with minimal changes, including the integration of cloned topic models and the extension to datasets from multiple institutions. This design ensures that ScholarNode can evolve beyond its initial MSU deployment and serve as a foundation for broader scholarly discovery and analysis.

8.2 System Architecture

The ScholarNode prototype was implemented as a lightweight web application that separates offline analytical processing from online interactive exploration. The goal of the architecture is to expose the topic-based collaboration framework to end users through a simple browser interface, while keeping the computationally expensive steps, such as topic modeling, community detection, and personalized ranking, outside of the request cycle. Figure 8.1 illustrates the main components.

8.2.1 Data Preparation and Integration Layer

All analytical steps described in Chapters 4–7 are executed offline to produce artifacts that the web application can consume. Topic models (LDA and BERT-based), researcher-level topic distributions, community assignments from NH-Louvain and Spectral–HAC, and recommendation scores are generated in batch and exported. During this stage, researcher identifiers (fid) and publication identifiers (workid) are preserved so that the web application

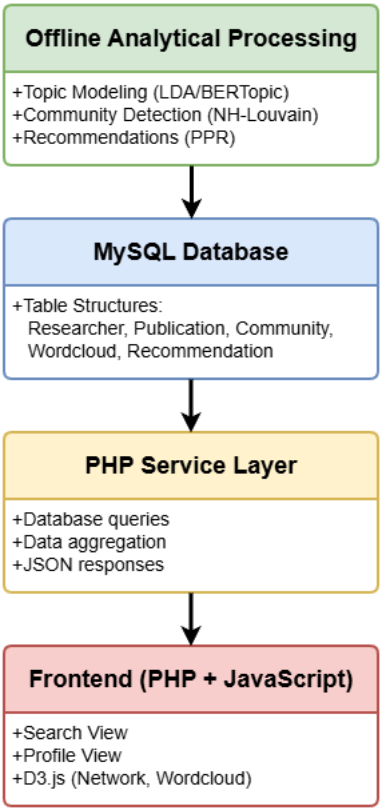


Figure 8.1: System architecture of the ScholarNode web application

can reference the exact entities used during experimentation.

The exported data are stored in a MySQL database that maintains the operational schema. The complete database entity–relationship diagram is provided in Appendix C. The database stores core relational information, including researchers, publications, researcher–publication mappings, coauthor mappings, and community memberships. In addition, two auxiliary tables are maintained for interpretability, an individual wordcloud table that stores, for each researcher, word and probability pairs representing their dominant topics, and a community wordcloud table that stores, for each community identifier, word and probability pairs representing the shared topical focus. A separate PageRank table is used to store the top ranked recommendations for each researcher, identified by fid, along with the recommended researcher identifier and the corresponding personalized PageRank score.

By materializing these results in MySQL, the web application can return recommendations without recomputing graph scores during user interaction.

8.2.2 Analytical and Recommendation Layer

The analytical logic in the prototype is kept minimal because the heavy computation is already reflected in the database tables. For a given researcher, the layer only needs to retrieve the precomputed outputs and assemble them into a coherent response. The PageRank table provides the top ten recommended collaborators and their scores, the community table provides the researcher's assigned community, and the wordcloud tables store the textual explanations at both individual and community levels. Although the current prototype does not incorporate cloned topic models, the database schema is designed to accommodate them. In a future extension, cloned entries would be stored with distinct identifiers linked to the original researcher, ensuring consistency with the modeling approach described in the earlier chapters.

This layer follows the same principle as the experimental pipeline. The web application consumes model outputs, it does not produce them. This decision keeps the online component responsive and also ensures that the results shown to users are exactly those evaluated in Chapter 7.

8.2.3 Service Layer in PHP

To connect the database to the browser, ScholarNode uses a simple service layer written in raw PHP. Although no web framework was used, the PHP scripts serve the role of an API by receiving HTTP requests, connecting to the MySQL database, running parameterized queries, and returning structured data. Typical endpoints include fetching a researcher profile by fid, searching researchers by name or by keyword, fetching the community identifier and corresponding JSON network file, retrieving the top 10 recommendations and their Personalized PageRank scores, and retrieving wordcloud terms.

Each script aggregates data from multiple tables, for example, a profile request will read from the researcher table (name, department, college, email), from the researcher–publication table (to show publication history), from the community table (to obtain the community id), from the PageRank table (to list recommended collaborators), and from the wordcloud tables (to provide interpretive terms). Returning a single structured response at the PHP layer keeps the frontend logic simple and reduces the number of database round trips.

8.2.4 Presentation and Visualization Layer

The frontend is also implemented in PHP, with embedded JavaScript to provide interactivity. Visualization is handled with d3.min.js [12] for the force-directed community view and with d3.layout.cloud [20] for rendering wordclouds. Community structures are saved as JSON files on the server, indexed by community or by researcher, and are loaded dynamically when a user opens a profile page. This design avoids computing graph layouts in the browser, since the JSON already contains the node and edge information produced offline.

The interface exposes three main views, a search view for name or concept lookup, a researcher profile view, and a topical or community browse view. In the profile view, the D3 force-directed graph shows the focal researcher and the surrounding community, with nodes colored by community membership and sized by degree or topical centrality. Clicking a node loads the corresponding researcher profile via PHP, which issues another database query. Adjacent panels display the individual wordcloud for the focal researcher and the community-level wordcloud for the detected community, both rendered from the word, probability pairs stored in MySQL. A recommendation panel lists the top 10 collaborators together with their PageRank scores and provides links to open their profiles.

Because most computations are done in advance, the browser is responsible only for rendering, filtering, and navigation. This makes it feasible to deploy the application on the

school server without requiring high-performance hardware.

8.2.5 Data Storage and Deployment

The application is deployed on an institutional server running a standard LAMP-style stack, with PHP handling requests and MySQL storing the operational data. Precomputed community JSON files are stored on disk and are referenced by the database so that each researcher profile can be associated with the correct network visualization. This hybrid storage, database for structured entities and file system for graph JSON, keeps the schema clean while still supporting large community structures.

Since all entities use stable researcher identifiers, the system can be extended to additional institutions by loading new researcher and publication records into MySQL, regenerating the community JSON files for the enlarged network, and adding the corresponding wordcloud entries. Because the user-facing layer is already written to consume generic fid and community identifiers, no major changes to the interface are required when scaling to other institutions (e.g., WSU and CSU).

8.2.6 Data Refresh and Maintenance

ScholarNode maintains an updated view of the scholarly network by periodically synchronizing with external publication data sources. The publication metadata, author affiliations, and coauthorship information are obtained from the OpenAlex API, which serves as the authoritative source for new works and institutional records. A Linux-based cron job executes automated scripts to check for new or modified entries on a scheduled basis, retrieving additional publications and newly added faculty profiles.

After new data are collected, the full analytical pipeline is rerun on the updated corpus. This includes topic modeling, community detection, wordcloud generation, and recomputation of the Personalized PageRank scores. The refreshed outputs are stored

locally and loaded back into the MySQL database, updating the community-level wordclouds, researcher-level topical summaries, and the recommendation tables.

The refresh interval is chosen based on how quickly new publications appear. Since research output at MSU grows gradually, updating the system every three months is enough to keep the topic models and recommendations current without rerunning the full pipeline too often. This schedule can be adjusted later depending on the amount of new data or the size of the institution. The goal is to keep ScholarNode up to date while keeping the processing workload manageable.

8.3 User Interface and Features

The ScholarNode web interface was designed to provide an interpretable and interactive environment for exploring scholarly communities and potential collaboration opportunities. The interface connects the analytical components described in earlier chapters to a user-friendly visualization platform, allowing users to search by concept, navigate researcher profiles, and interpret topical relationships through wordclouds. This section illustrates the major features of the system using three example views corresponding to the search, profile, and interpretability stages of the user workflow.

8.3.1 Search Interface

The search interface serves as the entry point for users to discover researchers by topic or by name. Users can issue keyword-based queries such as “Lidar,” “machine learning,” or “forest fire,” which are matched against indexed publication titles, abstracts, and topic distributions. The resulting page lists researchers associated with the query, showing their names, institutional affiliation, and publication count. For the dissertation figures shown here, researcher names are replaced with anonymized identifiers (for example, R-74), but the actual prototype displays full researcher names. Each search result links directly to

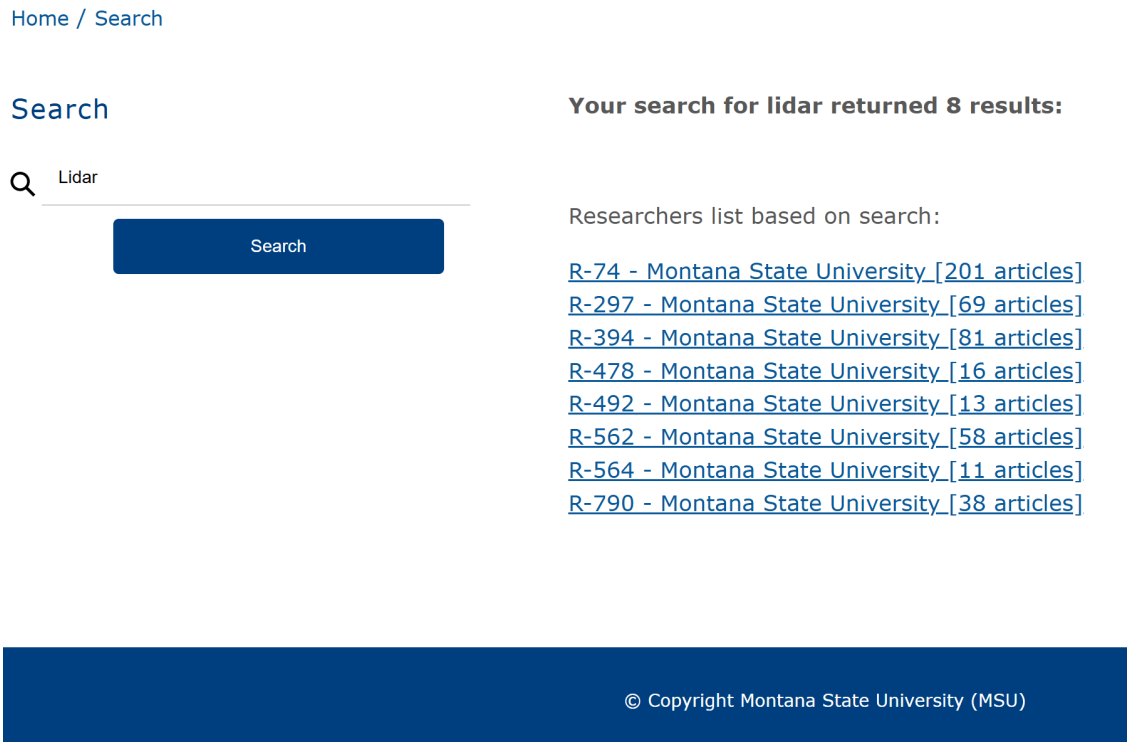


Figure 8.2: Search interface in ScholarNode showing example results for the keyword "Lidar"

the corresponding profile view, where community visualizations and recommendations are shown. Figure 8.2 illustrates the interface for the query "Lidar," where the system returns multiple relevant researchers ordered by topical similarity.

8.3.2 Researcher Profile View

Selecting a researcher from the search results opens the profile view, which visualizes the topical and collaboration context of that researcher within the broader scholarly network. Each node in the visualization represents a researcher, and node color indicates departmental affiliation, enabling users to visually assess the disciplinary diversity of a local network. Figure 8.3 shows the example profile view for R-74. The surrounding nodes represent researchers from multiple departments including Electrical Engineering, Civil Engineering, Land Resources and Environmental Science, and Physics, illustrating the interdisciplinary

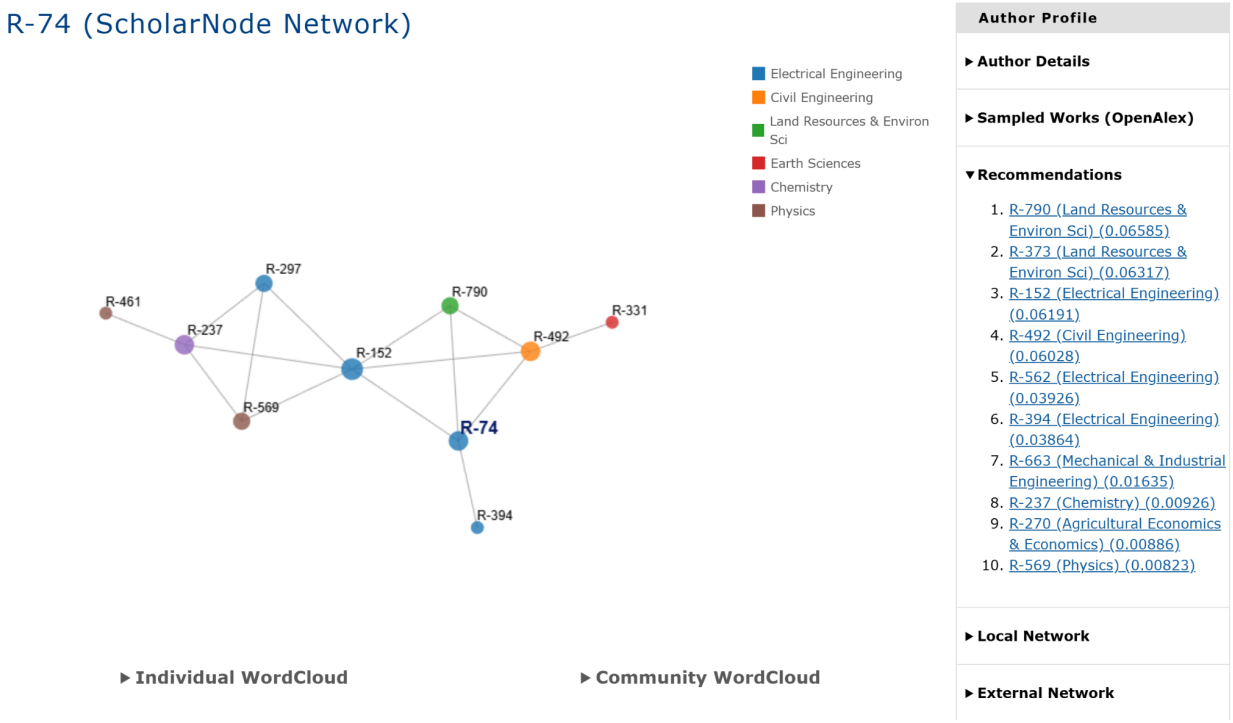


Figure 8.3: Researcher profile view in ScholarNode showing the interdisciplinary network

structure of the detected community. The right-hand panel lists the top recommended collaborators for R-74 along with their departments and Personalized PageRank scores. This panel provides an interpretable ranking of candidates based on both network structure and topical similarity. Expandable sections below the recommendations allow users to access the researcher's detailed information, publication samples, and topical wordclouds. Hovering over a node displays a tooltip with basic researcher details, including name, department, and total publication count. Both nodes and recommendation links are clickable, allowing direct navigation to the corresponding researcher's profile within the same network.

8.3.3 Individual and Community WordClouds

To further support interpretability, ScholarNode provides topical wordclouds that summarize the dominant research themes of both the focal researcher and the researcher's

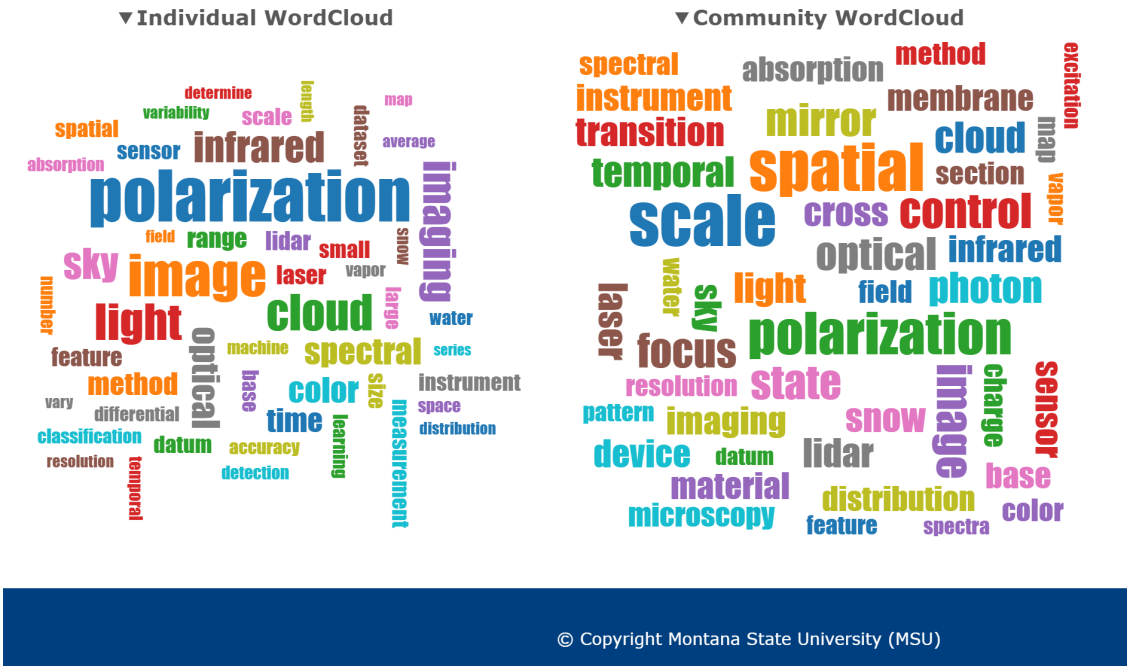


Figure 8.4: Individual and community-level wordclouds in ScholarNode showing dominant topical themes

community. The *Individual WordCloud* visualizes the highest-weighted terms in the researcher's topic distribution, while the *Community WordCloud* aggregates the top terms across all members of the community. As shown in Figure 8.4, the individual and community clouds for R-74 reveal overlapping and complementary terms such as "polarization", "image", and "spatial", highlighting topical coherence between the researcher and the surrounding community. These visualizations provide intuitive explanations for why certain collaborators are recommended, linking similarity scores to explicit topical content rather than opaque numerical metrics. Each keyword in the wordcloud is clickable, triggering a search for all researchers associated with that concept. This feature enables contextual exploration of topical clusters directly from the visualized terms.

Together, these three components illustrate the integrated workflow supported by ScholarNode. Users begin with concept-based exploration through the search interface,

navigate to the profile view to examine interdisciplinary connections and collaboration recommendations, and interpret the topical overlap between individual and community research through the wordclouds. By unifying these elements, ScholarNode demonstrates how topic-based modeling, community detection, and recommendation results can be presented in a transparent and interpretable manner, bridging analytical modeling with practical usability.

8.4 Summary

This chapter presented the ScholarNode prototype, a web-based system that operationalizes the topic-based collaboration framework developed in the previous chapters. The system integrates topic modeling, community detection, and recommendation mechanisms into an interactive platform that enables users to explore researchers, visualize interdisciplinary networks, and interpret recommendations through topical wordclouds. Designed initially for Montana State University, the prototype demonstrates how analytical modeling can be transformed into a practical and interpretable tool for collaboration discovery.

The architecture supports modular expansion, allowing integration of additional datasets from other institutions such as WSU and CSU. Its design also accommodates future incorporation of cloned model variants and the explainability components introduced in Chapter 7, such as stability metrics and topical overlap visualizations. Together, these extensions will enable broader deployment and more transparent recommendation behavior across institutional boundaries. ScholarNode therefore serves as a practical foundation linking topic-based modeling with user-oriented exploration and sets the stage for future large-scale, multi-institutional implementations.

CHAPTER NINE

CONCLUSION

This dissertation set out to advance content-based recommender systems by integrating topic modeling and social network analysis within a unified, interpretable framework. Traditional content-based approaches are effective in identifying item-level similarities but often ignore the relational structures that shape how knowledge, expertise, or creativity circulate within communities. By representing content similarity as a network, this study introduced a structured way to capture both the semantic and relational dimensions of data, leading to recommendations that are not only accurate but also explainable. The overarching goal was to improve interpretability, scalability, and balance in recommendation systems, ensuring that connections are discovered through meaningful patterns of content similarity and community context. To address these goals, the dissertation investigated four interrelated research questions, each contributing a distinct methodological and empirical advancement. The key findings and contributions are summarized below.

9.1 Summary of Contribution

RQ1: Topic model integration with social network analysis for interdisciplinary collaboration. The first research question investigated how topic modeling could be integrated with social network analysis to build topic-based collaboration networks that promote cross-domain recommendations. This work established the foundation of the framework by demonstrating how textual content, when transformed into topic distributions, can define weighted edges between entities based on their topical similarity. Using probabilistic LDA topic modeling, topic vectors were generated for each document and aggregated at the author level to capture individual topical profiles. These profiles enabled

the construction of fully connected networks where edge weights corresponded to the pairwise similarity between author-level topic distributions.

By systematically varying the similarity threshold, the resulting graphs revealed clear structural patterns that aligned with disciplinary boundaries while also exposing cross-domain links that coauthorship networks overlook. Applying community detection algorithms such as Louvain and Spectral Clustering allowed these networks to be partitioned into meaningful clusters representing latent research themes. Experiments across multiple institutions showed that moderate edge thresholds improved modularity and interpretability, leading to networks that balanced cohesion and diversity. These results confirmed that topic-based similarity provides a robust foundation for uncovering collaboration potential beyond visible ties. Portions of this work were presented in [86], which introduced the early version of the topic-based framework and established its applicability for interdisciplinary recommendations. This approach is generalizable beyond scholarly data and can be applied to any domain where items or users are characterized by textual or feature-rich content.

RQ2: Hierarchical community modeling for multilevel network understanding. The second research question explored how hierarchical community detection could improve both the structural interpretation of networks and the precision of recommendations. Traditional implementations of Louvain identify a single, non-overlapping community assignment for each vertex, resulting in a partition that reflects one dominant level of structure. However, real-world scholarly networks often exhibit hierarchical organization, where broader research areas contain multiple layers of increasingly specialized subcommunities. To capture this multilevel structure, two complementary hierarchical extensions were developed: Nested Hierarchical Louvain (NH-Louvain) and Spectral Hierarchical Agglomerative Clustering (Spectral–HAC).

The NH-Louvain algorithm extended standard modularity-based detection by recursively applying Louvain to each community, uncovering finer-grained subcommunities until

stability criteria were met. Spectral–HAC, on the other hand, combined eigenvector decomposition with agglomerative clustering to generate a dendrogram that reflects similarity-based merging decisions. Both methods were evaluated using modularity and cophenetic correlation coefficients, demonstrating that hierarchical models capture deeper community structures and provide more organized views of the network than single-level clustering.

These results revealed that hierarchical community detection not only captures structure at multiple levels but also enriches the recommendation process. Multi-level groupings allow for recommendations at varying levels of abstraction, ranging from tightly related collaborators within a subcommunity to broader interdisciplinary partners spanning higher-level clusters. This hierarchical representation supports multi-resolution recommendation, an aspect rarely explored in existing literature. The conceptual and empirical results of this component extended the topic-based framework introduced in [88].

RQ3: Mitigating content-data imbalance through cloning-based strategies. The third research question addressed an important yet underexplored challenge in content-based recommendation: data imbalance. In most real-world datasets, a small group of prolific entities dominates content generation, producing skewed topic distributions that bias both clustering and recommendation outcomes. This imbalance tends to obscure secondary or interdisciplinary interests, causing prolific individuals to appear in a single dominant topic space while their diverse research directions remain hidden.

To mitigate this, the dissertation introduced a cloning-based approach where prolific authors were represented by multiple topical instances, each reflecting a distinct theme derived from their local topic distributions. Two variants, Clone-LDA and Clone-BERT, were implemented, corresponding to the two topic modeling paradigms used in this study. Cloning expanded the representation space, enabling community detection algorithms to assign different topical subprofiles of the same researcher to appropriate communities. This resulted in networks that were more balanced and thematically diverse, with improved

representation of interdisciplinary edges.

Empirical analysis showed that cloning improved the recovery of held-out coauthorship links compared with non-cloned models and increased the diversity of recommendations by highlighting researchers' secondary topical areas. In particular, Clone-BERT produced finer separation of subtopics within prolific authors, while Clone-LDA retained stronger coherence at the thematic level. Together, these findings confirmed that cloning is an effective strategy for addressing representation skew in content-based networks.As detailed in the conference version of this study [89], addressing imbalance directly within the modeling process improves the representational balance of topic-based networks and enhances the diversity and coverage of the resulting recommendations.

RQ4: Evaluation, explainability, and system implementation. The fourth research question focused on evaluating the effectiveness and reliability of the proposed framework and on translating it into a functional system. Evaluation involved two complementary dimensions: accuracy and stability. Accuracy was assessed by measuring how well the system reproduced existing coauthorships and predicted potential collaborations using topic similarity as the predictive feature. Stability experiments simulated incomplete information by withholding subsets of coauthored publications during training, testing whether recommendations remained consistent under data perturbation. Results demonstrated that both LDA- and BERT-based models produced stable similarity scores even when 50% of the publication data was removed, confirming robustness to missing information.

Explainability was a central design goal throughout this work. Each recommendation in the framework can be traced to interpretable evidence in the form of shared topical weights and community co-membership. These explanations support user trust and adoption by helping researchers understand the rationale behind recommended collaborators and topics. To operationalize these principles, the ScholarNode web prototype was developed

as a practical demonstration of the framework's capabilities. The system integrates topic modeling, network construction, and visualization into an interactive interface that displays researcher profiles, community word clouds, and top-ranked recommendations. It provides both textual and visual explanations that link recommendations to their topical foundations. The ScholarNode prototype, initially introduced in [87], illustrates how the methodological advances of this dissertation can be implemented in a real-world, interpretable recommender system.

9.2 Limitations and Future Work

While this dissertation establishes a strong foundation for interpretable, network-based recommender systems, several limitations remain that inform opportunities for future refinement.

First, the analyses rely on publication metadata obtained through public APIs (i.e., OpenAlex). Although this choice aligns with the goal of evaluating interdisciplinarity through textual content rather than direct social connections, the completeness and accuracy of publication records cannot be guaranteed. Name inconsistencies, missing abstracts, and incomplete metadata occasionally affect the quality of topic representations and the resulting similarity computations. These data limitations introduce a degree of uncertainty that future work could mitigate through improved data reconciliation and metadata validation pipelines.

Second, topic modeling and network construction are computationally demanding processes. The similarity computation between all pairs of entities requires $O(n^2)$ operations, and hierarchical community detection further increases complexity. While these requirements remained manageable for the three-institution and combined datasets analyzed here, scaling the framework to larger multi-institutional or national corpora would require distributed or approximate methods to reduce computational cost. Additionally, topic interpretability and coherence depend on preprocessing quality and parameter tuning; although coherence

measures guided model selection, the subjective nature of topic interpretation and threshold selection introduces inherent modeling variability.

Third, the evaluation design uses historical coauthorship as a proxy for ground-truth collaboration, a common but imperfect assumption. Coauthorship captures realized collaborations but provides limited evidence for potential or novel ones. As a result, the evaluation metrics primarily measure the model's ability to reproduce existing patterns rather than its capacity to recommend diverse or previously unrealized collaborations. Developing evaluation protocols that explicitly account for novelty and diversity remains an open methodological challenge for future research.

Fourth, the current ScholarNode prototype focuses on recommendation generation and visualization but does not yet include mechanisms for user feedback, adaptive learning, or longitudinal tracking of user interaction. While the system provides interpretability through topic- and community-level explanations, future extensions could incorporate interactive feedback loops that allow users to refine recommendations and observe how their feedback influences underlying model behavior. Such functionality would not only improve personalization but also strengthen trust and transparency in practical deployments.

Finally, the present framework models textual content as the primary information source. Although this focus effectively captures semantic similarity, real-world recommender systems increasingly operate in multimodal environments where content may include images, videos, or sensor signals. Adapting the framework to process and integrate such heterogeneous content types could broaden its applicability while maintaining the same topic-based and network-oriented reasoning principles established in this study.

A particularly promising direction for future research involves extending the current static framework into a dynamic or temporal modeling setting. In its present form, the network represents a snapshot of scholarly content aggregated over time, assuming that topical similarity and community structures remain stable. However, real-world research

and content ecosystems evolve continuously as new publications, authors, and themes emerge. Incorporating temporal dynamics would allow the framework to model how topic distributions, similarity weights, and community memberships change over time. Such temporal extensions could leverage incremental topic modeling, streaming embeddings, or dynamic graph algorithms to update the network efficiently without full recomputation. This capability would enable the system to capture emerging interdisciplinary areas, track the evolution of research communities, and provide time-aware recommendations that reflect current trends and shifting expertise landscapes.

In addition to these methodological extensions, the framework's design and underlying principles are broadly transferable to other recommendation settings. The same principles, representing content similarity as a network, detecting hierarchical communities, mitigating imbalance, and maintaining explainability, apply to organizational knowledge management, interdisciplinary team formation, research–industry matchmaking, or educational content discovery. Any domain where content reflects expertise, intent, or thematic focus can benefit from this framework's ability to balance interpretability with scalability. By uniting semantic representation and network reasoning, this research provides a foundation for a new generation of explainable content-based recommender systems that generalize across domains and foster more meaningful, equitable connections.

At a theoretical level, this work contributes to the understanding of how topic-based representations can be embedded into network structures to capture higher-order semantic relationships, bridging natural language processing and network science perspectives on recommender design.

In summary, this dissertation demonstrates that integrating topic modeling with network analysis produces recommender systems that maintain consistent performance under partial information and offer clearer, more contextually organized recommendations. By addressing the challenges of structural modeling, data imbalance, and interpretability, this

research contributes new theoretical insights and practical methods that extend far beyond the academic domain. The ScholarNode system exemplifies how these concepts can be implemented in a real application, offering a concrete step toward the broader goal of creating recommendation systems that predict relevance and clearly explain the reasoning behind it.

REFERENCES CITED

[1] Zeinab Abbassi, Sihem Amer-Yahia, Laks V.S. Lakshmanan, Sergei Vassilvitskii, and Cong Yu. Getting recommender systems to think outside the box. In *Proceedings of the Third ACM Conference on Recommender Systems*, page 285–288, 2009.

[2] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[3] Reid Andersen, Fan R.K. Chung, and Kevin J. Lang. Local graph partitioning using PageRank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, 2006.

[4] Lars Backstrom and Jure Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 635–644, 2011.

[5] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[6] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, 35(12):29–38, 1992.

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, 2019.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[9] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[10] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.

[11] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.

[12] Mike Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3.js: Data-driven documents. `https://d3js.org/`, 2011. Accessed: 2025-10-31.

[13] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[14] Raymond B Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[15] Mohamed Amine Chatti, Mouadh Guesmi, and Arham Muslim. Visualization for recommendation explainability: A survey and new perspectives. *ACM Transactions on Interactive Intelligent Systems*, 14(3):1–40, 2024.

[16] Petr Chunaev. Community detection in node-attributed social networks: A survey. *Computer Science Review*, 37:100286, 2020.

[17] Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[18] Aaron Clauset, Mark E.J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 70(6):066111, 2004.

[19] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[20] Jason Davies. d3-cloud: Word cloud layout for d3.js. `https://github.com/jasondavies/d3-cloud`, 2013. Accessed: 2025-10-31.

[21] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977.

[22] Mitali Desai, Rupa G Mehta, and Dipti P Rana. Scholarrec: A scholars' recommender system that combines scholastic influence and social collaborations in academic social networks. *International Journal of Data Science and Analytics*, 16(2):203–216, 2023.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, 2019.

[24] Daizong Ding, Mi Zhang, Hanrui Wang, Xudong Pan, Min Yang, and Xiangnan He. A deep learning framework for self-evolving hierarchical community detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 372–381, 2021.

[25] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[26] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16 (1-2):1–177, 2022.

[27] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891, 2022.

[28] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[29] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[30] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40 (1):35–41, 1977.

[31] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

[32] Yingqiang Ge, Shuchang Liu, Zuohui Fu, Juntao Tan, Zelong Li, Shuyuan Xu, Yunqi Li, Yikun Xian, and Yongfeng Zhang. A survey on trustworthy recommender systems. *ACM Transactions on Recommender Systems*, 3(2):1–68, 2024.

[33] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 6, pages 721–741, 1984.

[34] Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems*, 16, 2003.

[35] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[36] Mouadh Guesmi, Mohamed Amine Chatti, Shoeb Joarder, Qurat Ul Ain, Clara Siepmann, Hoda Ghanbarzadeh, and Rawaa Alatrash. Justification vs. transparency: Why and how visual explanations in a scientific literature recommender system. *Information*, 14(7):401, 2023.

[37] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

[38] Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[39] Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954.

[40] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[41] Myriam Hernández-Alvarez and José M Gomez. Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, 22(3):327–349, 2016.

[42] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[43] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*, volume 4, pages 9–56, 2008.

[44] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[45] Shi Hui, Ma Wei, Zhang XiaoLiang, Jiang JunYan, Liu YanBing, and Chen ShuJuan. A hybrid paper recommendation method by using heterogeneous graph and metadata. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[46] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[47] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, 2010.

[48] Chaker Jebari, Enrique Herrera-Viedma, and Manuel Jesus Cobo. The use of citation context to detect the evolution of research topics: a large-scale analysis. *Scientometrics*, 126(4):2971–2989, 2021.

[49] Di Jin, Zhizhi Yu, Pengfei Jiao, Shirui Pan, Dongxiao He, Jia Wu, Philip S Yu, and Weixiong Zhang. A survey of community detection approaches: From statistical modeling to deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1149–1170, 2021.

[50] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37: 183–233, 1999.

[51] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, NJ, 2008.

[52] Subhash Khot. The hardness of graph partitioning. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 544–553, 2004.

[53] Orrin Edgar Klapp. *Overload and Boredom: Essays on the Quality of Life in the Information Society*. Greenwood Publishing Group Inc., 1986.

[54] Jon M. Kleinberg. On the complexity of the graph clustering problem. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 918–926, 2003.

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

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

[57] Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

[58] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[59] Sameer Kumar. Co-authorship networks: A review of the literature. *Aslib Journal of Information Management*, 67(1):55–73, 2015.

[60] Andrea Lancichinetti and Santo Fortunato. Limits of modularity maximization in community detection. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 84(6):066122, 2011.

[61] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[62] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636, 2006.

[63] Jing Li, Feng Xia, Wei Wang, Zhen Chen, Nana Yaw Asabere, and Huizhen Jiang. ACRec: A co-authorship based random walk model for academic collaboration recommendation. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1209–1214, 2014.

[64] Ruihuang Li, Changqing Zhang, Qinghua Hu, Pengfei Zhu, and Zheng Wang. Flexible multi-view representation learning for subspace clustering. In *International Joint Conference on Artificial Intelligence*, volume 2019, pages 2916–2922, 2019.

[65] Tianxi Li, Lihua Lei, Sharmodeep Bhattacharyya, Koen Van den Berge, Purnamrita Sarkar, Peter J Bickel, and Elizaveta Levina. Hierarchical community detection by recursive partitioning. *Journal of the American Statistical Association*, 117(538): 951–968, 2022.

[66] Weijuan Li. Scientific paper recommender system using deep learning and link prediction in citation network. *Heliyon*, 10(14), 2024.

[67] Huizhi Liang, Yue Xu, Dian Tjondronegoro, and Peter Christen. Time-aware topic recommendation based on micro-blogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1657–1661, 2012.

[68] Wei Liang, Xiaokang Zhou, Suzhen Huang, Chunhua Hu, and Qun Jin. Recommendation for cross-disciplinary collaboration based on potential research field discovery. In *Fifth International Conference on Advanced Cloud and Big Data (CBD)*, pages 349–354, 2017.

[69] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 556–559, 2003.

[70] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 2002.

[71] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications*, 390(6):1150–1170, 2011.

[72] X. Lu, L. Zhang, Y. Liu, and G. Zhang. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 47(4): 58, 2015.

[73] Hao Ma, Tom Chao Zhou, Michael R. Lyu, and Irwin King. Improving recommender systems by incorporating social contextual information. *ACM Transactions on Information Systems*, 29(2), 2011.

[74] Ilya Makarov and Olga Gerasimova. Predicting collaborations in co-authorship network. In *14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 1–6. IEEE, 2019.

[75] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[76] Frank J Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[77] Leland McInnes, John Healy, and Steve Astels. HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[78] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[79] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 262–272, 2011.

[80] Regina Moirano, Marisa Analía Sánchez, and Libor Štěpánek. Creative interdisciplinary collaboration: A systematic literature review. *Thinking Skills and Creativity*, 35:100626, 2020.

[81] Mark E.J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

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

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

[84] Mark E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[85] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. Spectral clustering for image segmentation. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pages 849–856, 2002.

[86] Md Asaduzzaman Noor, John Sheppard, and Jason Clark. Finding potential research collaborations from social networks derived from topic models. In *10th International Conference on Behavioural and Social Computing (BESC)*, pages 1–7. IEEE, 2023.

[87] Md Asaduzzaman Noor, Jason A Clark, and John W Sheppard. 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*, pages 2791–2795, 2024.

[88] Md Asaduzzaman Noor, John W Sheppard, and Jason A Clark. Identifying hierarchical community structures in content-based scholarly social networks. In *International Conference on Machine Learning and Applications (ICMLA)*, pages 440–447. IEEE, 2024.

[89] Md Asaduzzaman Noor, John Sheppard, and Jason Clark. Handling publication imbalance for effective community detection in scholarly networks. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1–8, 2025.

[90] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999. URL `http://ilpubs.stanford.edu:8090/422/`. Technical Report.

[91] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*, pages 325–341. Springer, 2007.

[92] Tiago P Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047, 2014.

[93] Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*, 2205.01833, 2022.

[94] Robert Clay Prim. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401, 1957.

[95] Diana Purwitasari, Chastine Fatichah, Surya Sumpeno, Christian Steglich, and Mauridhi Hery Purnomo. Identifying collaboration dynamics of bipartite author-topic networks with the influences of interest changes. *Scientometrics*, 122(3):1407–1443, 2020.

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

[97] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 74(1):016110, 2006.

[98] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[99] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[100] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender Systems Handbook*, pages 1–35. Springer, 2010.

[101] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, 2015.

[102] N. J. Roland, R. D. R. McRae, and A. W. McCombe. *Key Topics in Otolaryngology.* CRC Press, 2 edition, 2000.

[103] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494, 2004.

[104] Sheldon M. Ross. *Introduction to Probability Models.* Academic Press, San Diego, CA, USA, sixth edition, 1997.

[105] M. Rosvall and C. T. Bergstrom. Detecting hierarchical community structures in networks using Infomap. *Information Sciences*, 179(15):3080–3091, 2009.

[106] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[107] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[108] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[109] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007.

[110] Upendra Shardanand and Pattie Maes. Social information filtering: algorithms for automating "word of mouth". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 210–217, 1995.

[111] Philip Simon, Peter Bamidele Shola, and Ovye John Abari. Application of content-based approach in research paper recommendation system for a digital library. *International Journal of Advanced Computer Science and Applications*, 5(10), 2014.

[112] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001.

[113] Robert R. Sokal and F. James Rohlf. The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40, 1962.

[114] Nasim Sonboli, Robin Burke, Michael Ekstrand, and Rishabh Mehrotra. The multisided complexity of fairness in recommender systems. *AI magazine*, 43(2): 164–176, 2022.

[115] Karen Sparck Jones. *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, page 132–142. Taylor Graham Publishing, 1988.

[116] Daniel A. Spielman. Spectral graph theory and its applications. *Linear Algebra and its Applications*, 438(11):2415–2445, 2013.

[117] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(Dec): 583–617, 2003.

[118] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson, 2005.

[119] Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4):100729, 2023.

[120] Paola Velardi, Roberto Navigli, Alessandro Cucchiarelli, and Fulvio D'Antonio. A new content-based model for social network analysis. In *IEEE International Conference on Semantic Computing*, pages 18–25, 2008.

[121] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17 (4):395–416, 2007.

[122] Scott Allen Wahl. *Hierarchical fuzzy spectral clustering in campaign finance social networks*. PhD thesis, Montana State University-Bozeman, Gianforte School of Computing, 2021.

[123] Shoujin Wang, Xiuzhen Zhang, Yan Wang, and Francesco Ricci. Trustworthy recommender systems. *ACM Transactions on Intelligent Systems and Technology*, 15(4):1–20, 2024.

[124] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3): 1–43, 2023.

[125] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[126] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.

[127] Frank Wilcoxon. Individual comparisons by ranking methods. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer, 1992.

[128] Naijia Xiao, Aifen Zhou, Megan L. Kempher, Benjamin Y. Zhou, Zhou Jason Shi, Mengting Yuan, Xue Guo, Linwei Wu, Daliang Ning, Joy Van Nostrand, Mary K. Firestone, and Jizhong Zhou. Disentangling direct from indirect relationships in association networks. *Proceedings of the National Academy of Sciences*, 119: e2109995119, 2022.

[129] Yunhong Xu, Duanning Zhou, and Jian Ma. Scholar-friend recommendation in online academic communities: An approach based on heterogeneous network. *Decision Support Systems*, 119:1–13, 2019.

[130] Chen Yang, Jianshan Sun, Jian Ma, Shanshan Zhang, Gang Wang, and Zhongsheng Hua. Scientific collaborator recommendation in heterogeneous bibliographic networks. In *48th Hawaii International Conference on System Sciences*, pages 552–561, 2015.

[131] Donghui Yang, Mingyang Zhang, Zhaoyang Shi, and Kehui Zhu. Enhancing accuracy and diversity of recommendation for interdisciplinary journals: A deep learning approach. *Journal of Information Science*, page 01655515251353177, 2025.

[132] Yang Yang, Ryan N Lichtenwalter, and Nitesh V Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, 45(3):751–782, 2015.

[133] Diao Yu and Xue Hong. Research on the tf-idf literature recommendation based on behavior data of electronic resource: Taking off-campus access system of electronic resource as an example. *Library Journal*, 41(12):45, 2022.

[134] Tong Zhang, Xiaotong Zhang, Xia Zhang, Jing Zhang, and Rui Zhang. A survey on machine learning for social networks: Analysis and applications. *ACM Computing Surveys (CSUR)*, 54(3):1–38, 2021.

[135] Yongfeng Zhang and Xu Chen. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101, 2020.

[136] Jianan Zhao, Qianlong Wen, Shiyu Sun, Yanfang Ye, and Chuxu Zhang. Multi-view self-supervised heterogeneous graph embedding. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 319–334. Springer, 2021.

[137] Qinghai Zheng, Jihua Zhu, and Zhongyu Li. Collaborative unsupervised multi-view representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4202–4210, 2021.

[138] Xiaokang Zhou, Wei Liang, Kevin I-Kai Wang, Runhe Huang, and Qun Jin. 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.

APPENDICES

APPENDIX A

MODULARITY AND COMMUNITY STATISTICS

Table A.1: Louvain and Spectral community statistics on WSU across selected thresholds

TH	**Comm**	**AvgJ**	**Max(L)**	**Min(L)**	**Med(L)**	**Std(L)**
			Max(S)	**Min(S)**	**Med(S)**	**Std(S)**
0.00	4	0.89	232	101	140	48.50
			211	124	139	33.90
0.12	4	0.89	232	101	140	48.50
			211	125	138.5	33.80
0.36	6	0.51	148	2	112.5	46.83
			229	49	76.5	59.16
0.48	14	0.55	84	2	44.5	27.07
			201	4	33.5	46.00
0.55	28	0.69	56	4	21	14.21
			59	4	18.5	13.31
0.60	30	0.89	66	3	13	16.66
			65	3	15.5	16.31
0.65	39	0.89	83	2	6	20.40
			80	3	6	20.25
0.70	38	0.85	93	2	3.5	25.78
			95	2	4	26.44

This appendix includes the full threshold based modularity and community statistics for WSU, CSU, and the combined MWC network. These tables complement the results discussed in Chapter 4, where the MSU analysis is presented in the main text.

Table A.2: Louvain and Spectral community statistics on CSU across selected thresholds

TH	**Comm**	**AvgJ**	**Max(L)**	**Min(L)**	**Med(L)**	**Std(L)**
			Max(S)	**Min(S)**	**Med(S)**	**Std(S)**
0.00	3	0.85	273	229	243	18.35
			279	198	268	35.87
0.12	3	0.85	273	228	244	18.62
			278	199	268	35.12
0.36	6	0.63	151	80	128.5	23.23
			232	52	111	55.11
0.48	15	0.65	91	14	53	25.59
			190	14	32	43.61
0.55	30	0.73	83	3	20	19.01
			107	1	19	24.37
0.60	37	0.79	60	2	13	16.31
			65	4	17	14.23
0.65	40	0.88	69	3	13.5	15.68
			69	4	13	15.49
0.70	40	0.83	101	2	14	19.11
			112	3	13	19.73

Table A.3: Louvain and Spectral community statistics on the combined MWC dataset

TH	**Comm**	**AvgJ**	**Max(L)**	**Min(L)**	**Med(L)**	**Std(L)**
			Max(S)	**Min(S)**	**Med(S)**	**Std(S)**
0.00	3	0.92	616	435	583	78.71
			617	431	586	81.36
0.12	3	0.89	609	447	578	70.21
			599	447	588	69.21
0.36	10	0.57	295	36	160.5	86.06
			509	18	123.5	134.90
0.48	30	0.64	146	2	39.5	41.90
			278	10	41	58.34
0.55	65	0.87	160	2	14	32.20
			166	4	14	33.21
0.60	70	0.85	230	2	5.5	43.98
			241	2	5.5	44.95
0.65	77	0.84	264	2	3	48.00
			264	2	4	47.84
0.70	51	0.84	296	2	5	62.63
			299	2	6	61.81

APPENDIX B

ADDITIONAL HIERARCHICAL DENDROGRAM VISUALIZATIONS

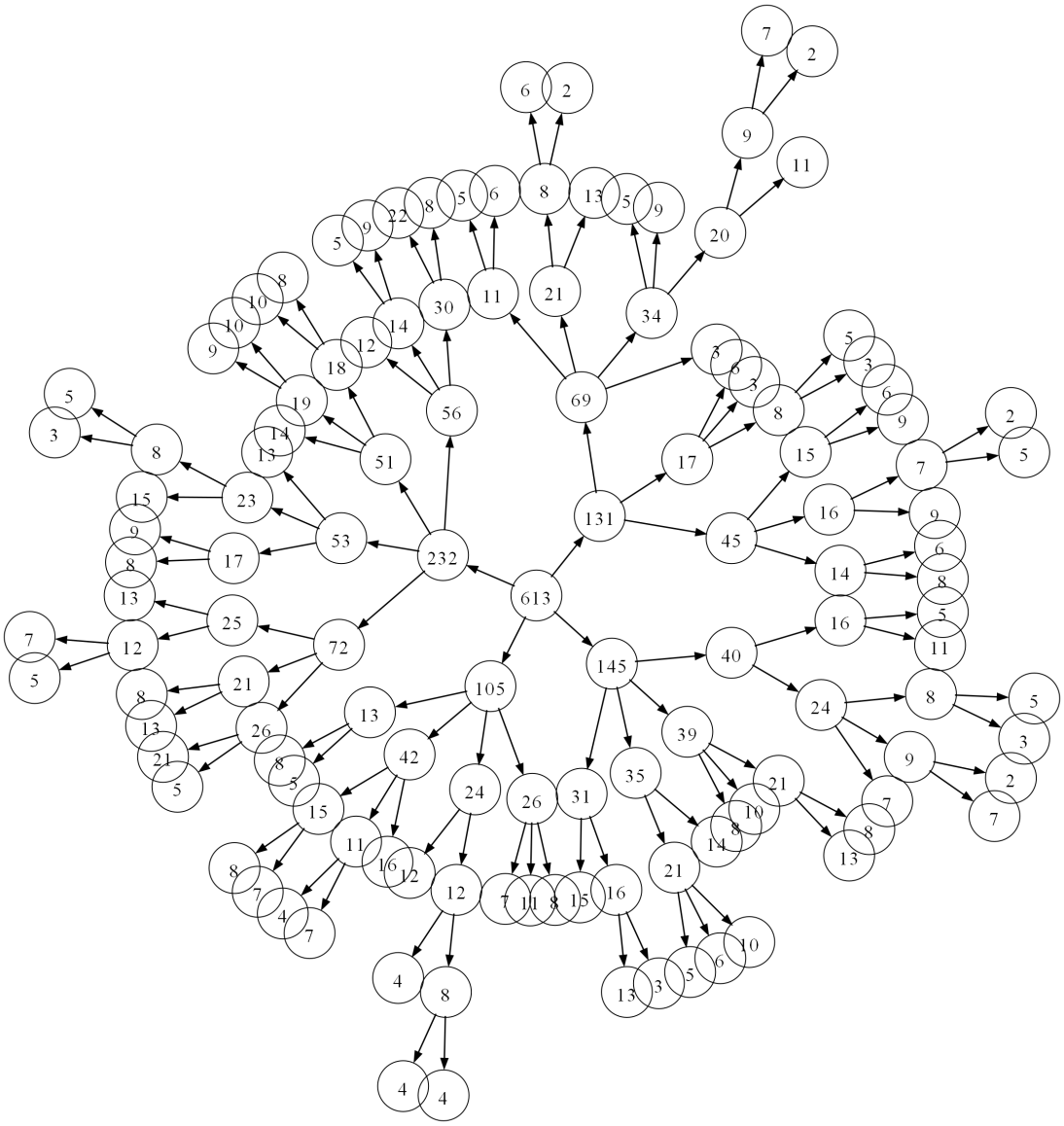
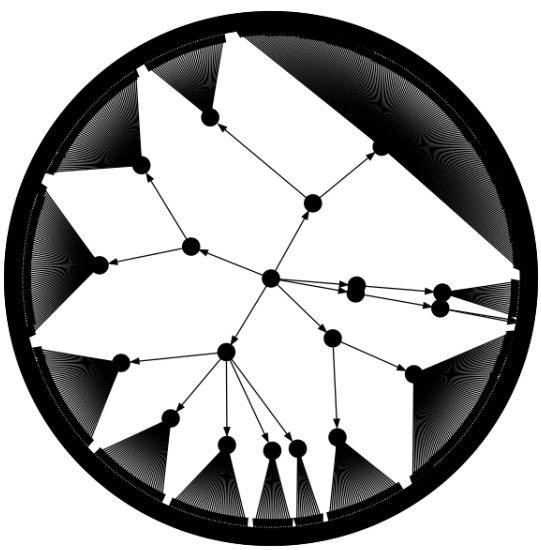
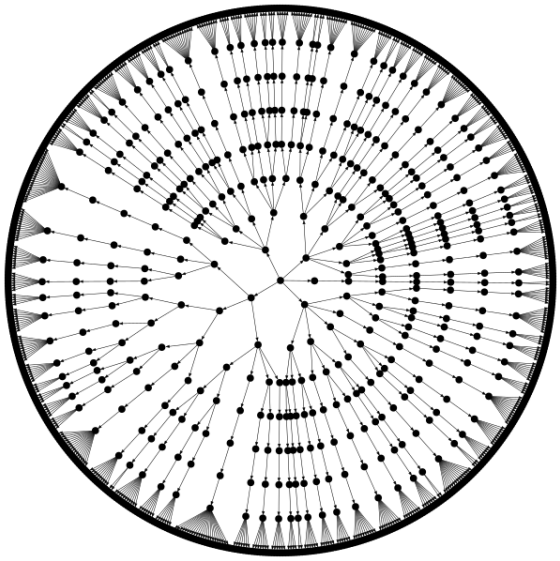


Figure B.1: NH-Louvain dendrogram of the WSU network

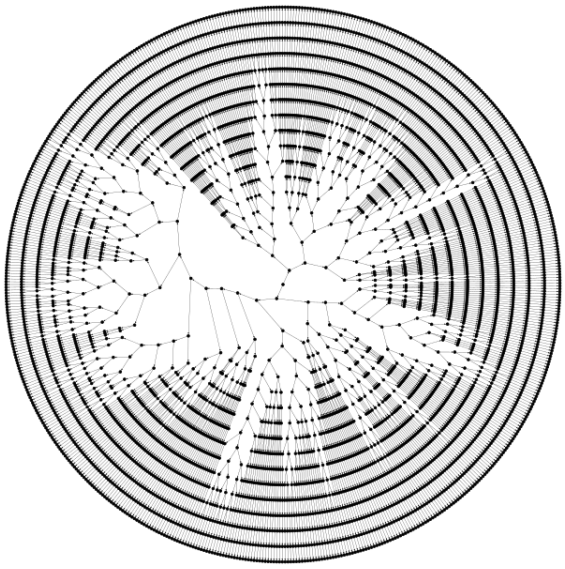
This appendix provides supplementary dendrogram visualizations for the WSU, CSU, and combined MWC networks. These figures complement the MSU examples discussed in Chapter 5, where we illustrated the hierarchical structure using MSU as a representative case. Although MSU offers a balanced and interpretable example for in-text discussion, the remaining institutions exhibit similar hierarchical patterns.



(a) Louvain



(b) NH-Louvain



(c) Spectral–HAC

Figure B.2: Dendrograms of the WSU-0.1 network across three hierarchical methods

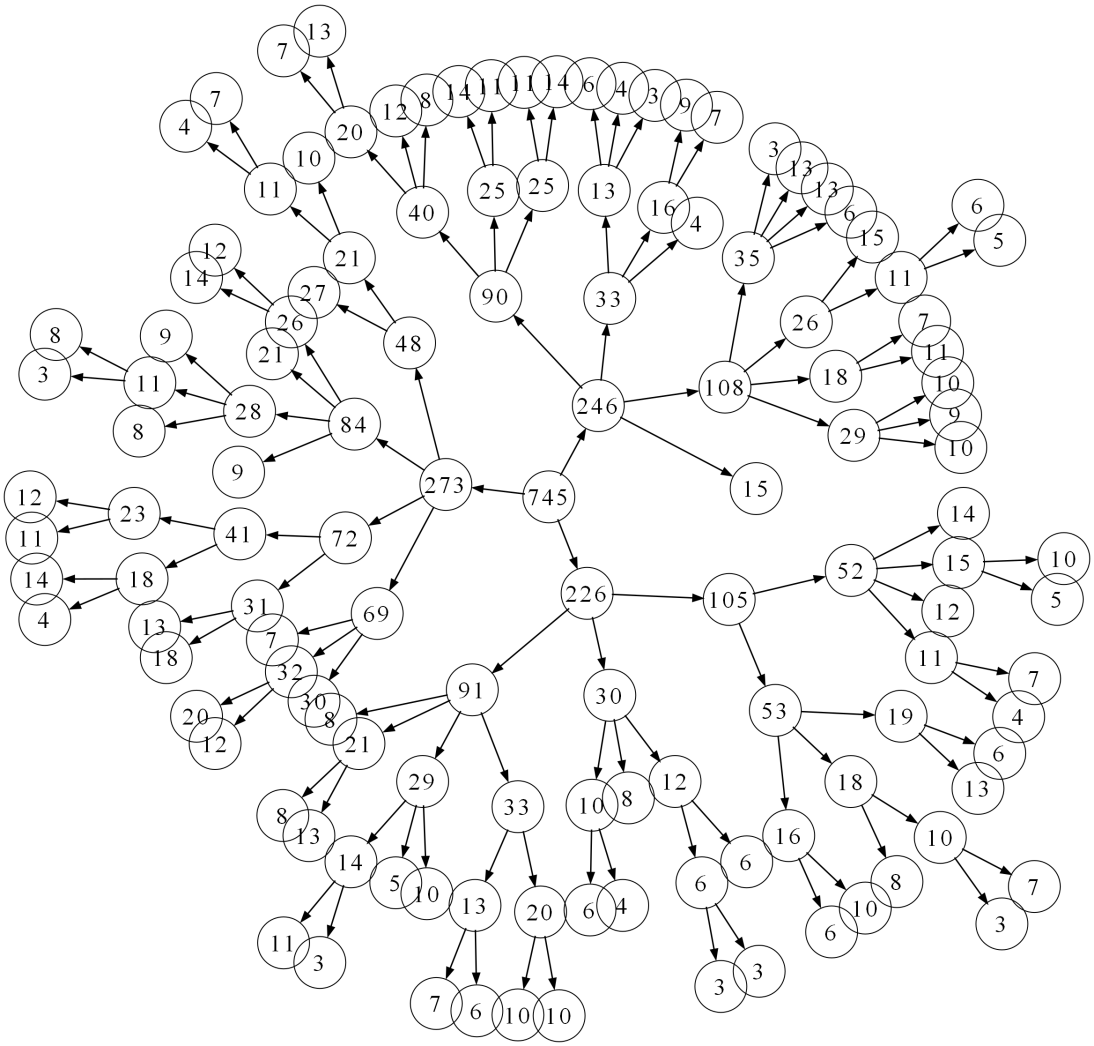
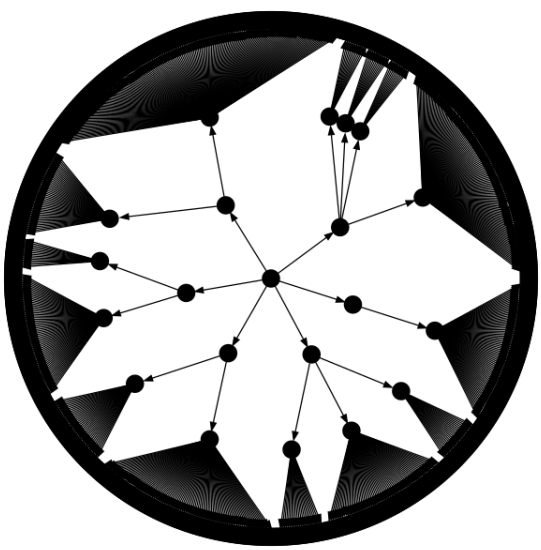
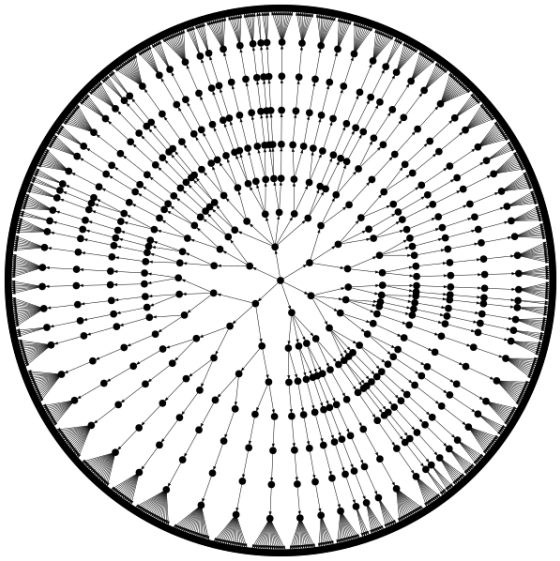


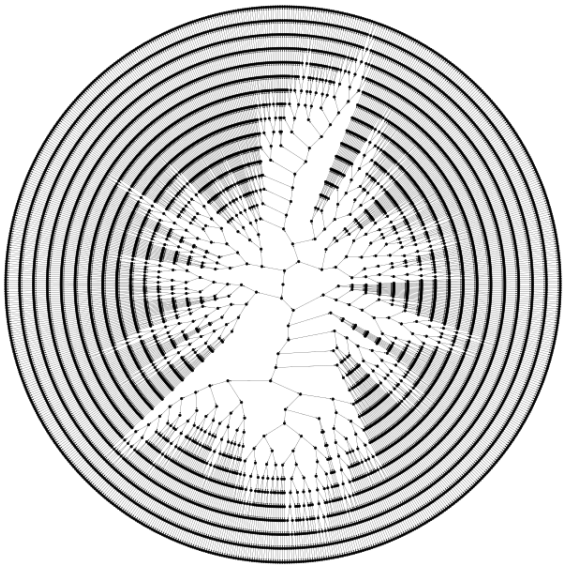
Figure B.3: NH-Louvain dendrogram of the CSU network



(a) Louvain



(b) NH-Louvain



(c) Spectral–HAC

Figure B.4: Dendrograms of the CSU-0.1 network across three hierarchical methods

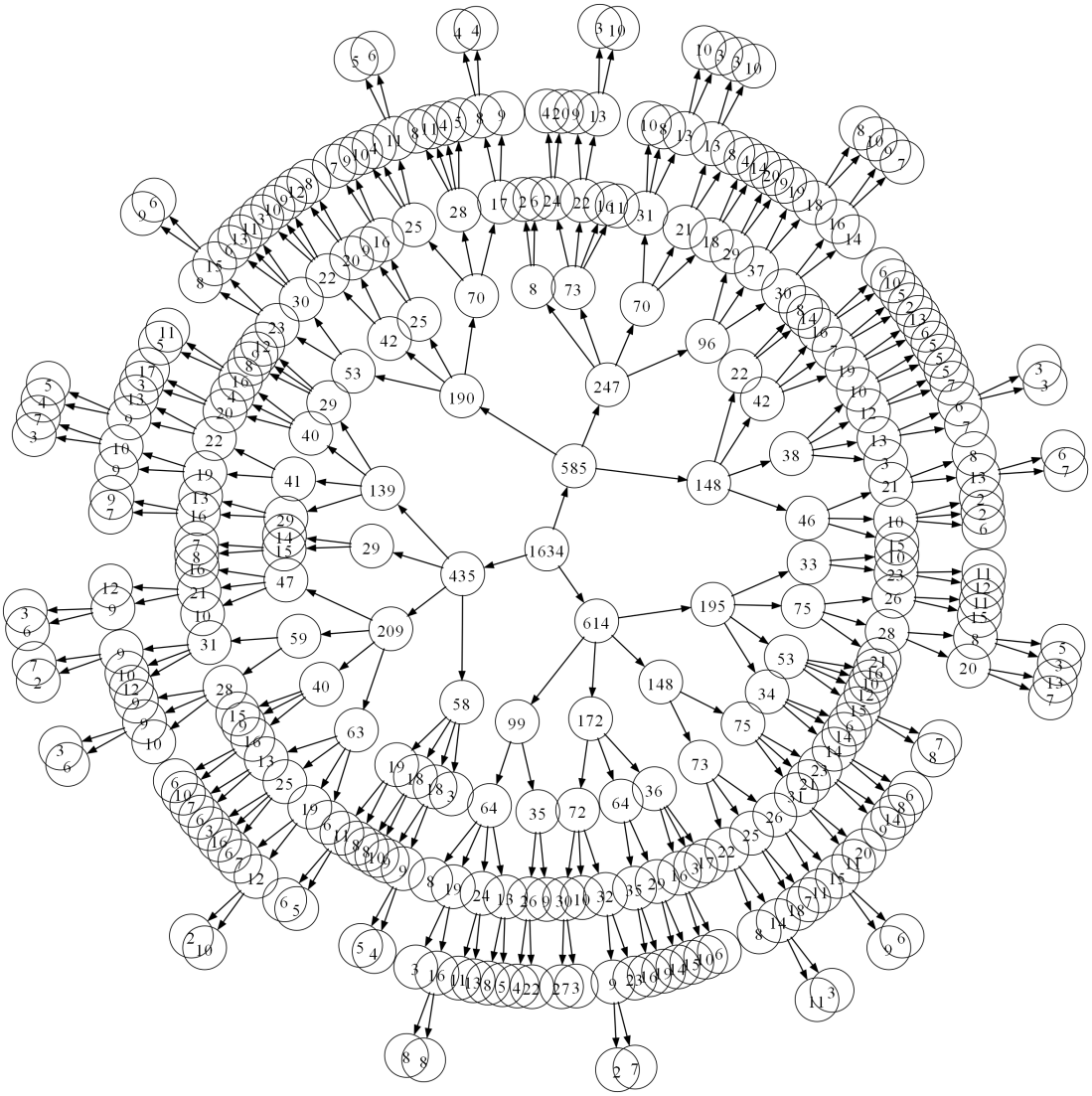
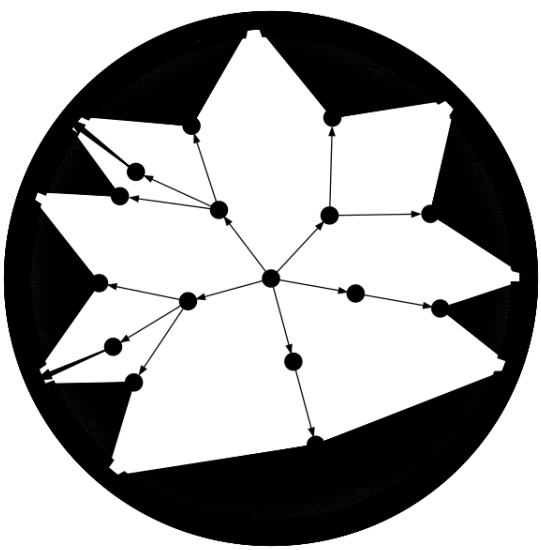
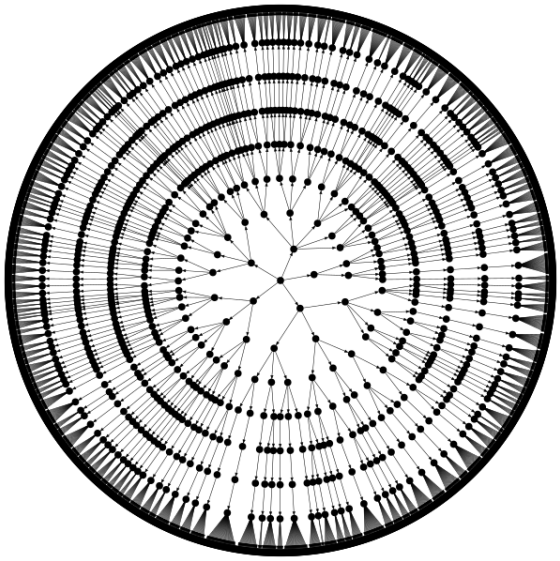


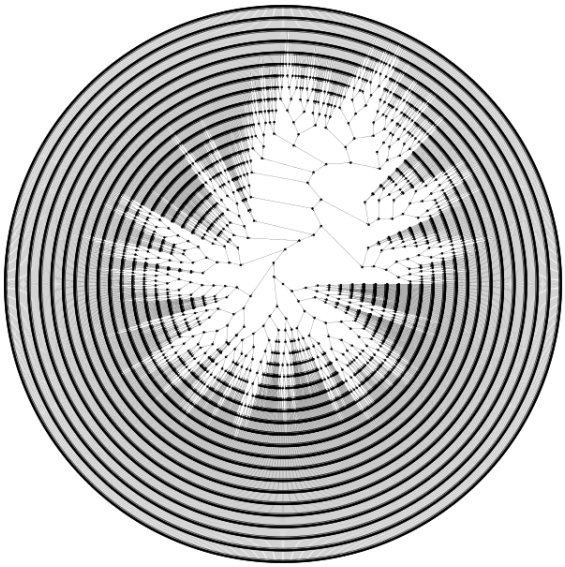
Figure B.5: NH-Louvain dendrogram of the combined MWC network



(a) Louvain



(b) NH-Louvain



(c) Spectral–HAC

Figure B.6: Dendrograms of the combined MWC-0.1 network across hierarchical methods

APPENDIX C

SCHOLARNODE ENTITY–RELATIONSHIP DIAGRAM

This appendix presents the entity–relationship (ER) diagram used in the ScholarNode prototype. It summarizes the core data structures supporting faculty records, OpenAlex author profiles, publication metadata, community assignments, PageRank scores, and term distributions. The diagram captures the relationships among the primary tables, including the `msu_faculty`, `authors`, `works`, and their associative mappings.

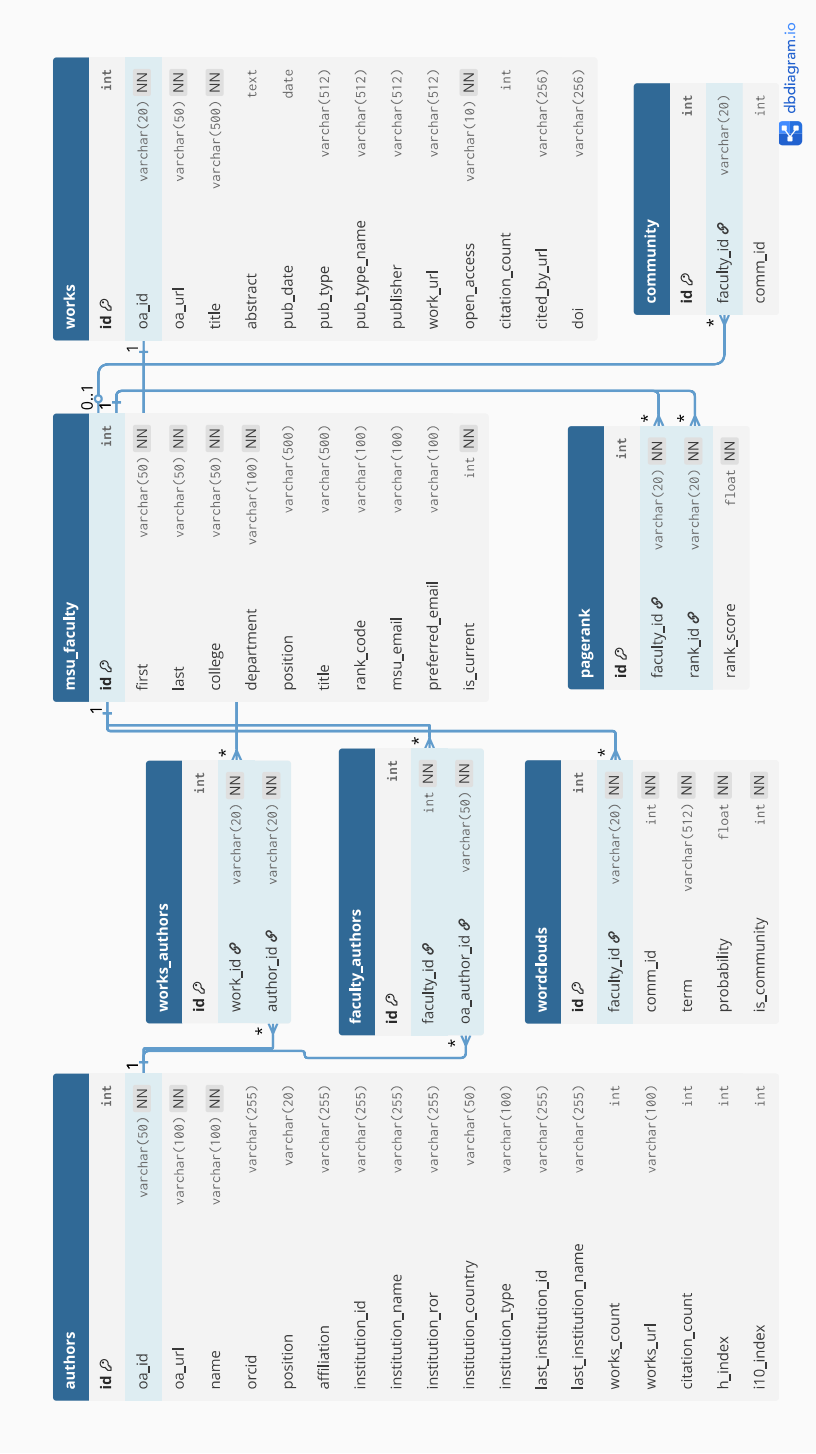


Figure C.1: Entity–relationship diagram of the ScholarNode prototype