Vrije Universiteit Amsterdam

Universiteit van Amsterdam

Master Thesis

Directing the force: Mapping the eScience Center's Network

Author: Zhining Bai (13714538)

1st supervisor: Adam Belloum
daily supervisor: Peter Kalverla (Netherlands eScience Center)
2nd reader: Rob van Nieuwpoort

A thesis submitted in fulfillment of the requirements for the joint UvA-VU Master of Science degree in Computer Science

July 20, 2023

"I am the master of my fate, I am the captain of my soul"
from Invictus*, by William Ernest Henley*

Abstract

As a consequence of the rapidly growing trend of multidisciplinary research collaborations and their potential to foster scientific innovation and research quality, understanding the dynamics of these collaborative networks is critical to the advancement of academic research.

The main goal of this study was to create a tool for searching and generating research collaboration networks for Dutch scientists, which could assist in understanding the intricacies of their research collaborations and provide actionable insights for identifying potential collaborative opportunities. Using a dataset of Dutch academic publications consisting of Scopus, Research Software Directory (RSD), and OpenAlex, we use Social Network Analysis (SNA) techniques to construct and examine the research collaboration networks.

We adopt an approach that explores networks from an individual researcher's perspective and analyzes research collaborations based on specific topics or subject areas. We have constructed an interactive network model, providing a clear overview of collaborative relationships among Dutch researchers and revealing key actors and communities within these networks. This study provides a transformative approach to understanding and strengthening research collaboration in the Netherlands by creating an intuitive, user-centered tool, which lays the groundwork for future academic collaboration.

Acknowledgements

First and foremost, I wish to express my sincere gratitude to my academic advisor, Prof.Dr. Adam Belloum. Your guidance, patience, and expertise were invaluable during my research journey. I am equally indebted to my daily supervisor, Dr. Peter Kalverla, for his invaluable advice and constant encouragement. The completion of this thesis would not have been possible without his consistent mentorship. I also wish to extend my thanks to all the members of Team Beta at the eScience Center, who contributed significantly to my project.

My time at the Netherlands eScience Center was unforgettable, and I am grateful for the opportunity to work in such an environment teeming with brilliant researchers. The presentations, daily chats, and exchange of ideas have broadened my perspective, and I have gained a wealth of knowledge that extends beyond academic research. I am grateful to all those who been involved in this journey.

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

Over recent years, collaboration between scientists becomes the norm, and it is widely known that good-quality research collaboration can enhance the productivity of individual scientists in many cases (5). Cross-disciplinary directions such as Artificial Intelligence (neural networks), Bioinformatics, and Nanoscience, as well as the advancement of traditional disciplines, promote multidisciplinary cross-collaboration. Collaboration among researchers has become increasingly important in the academic community due to its potential to promote scientific innovation and enhance research quality. Therefore, finding potential research collaborators and expanding collaboration networks is critical in the development of disciplinary research. (6) Building collaboration networks of researchers across the Netherlands, for researchers, it summarizes and analyzes researchers' collaboration history, provides a summary of their academic collaboration results, and allows analysis of historical collaboration information to predict researchers' collaboration tendencies. For academic research development, it provides an overview of academic collaborations in recent years, summarizes collaboration tendencies in different disciplines, and provides direction and guidance for future academic collaborations, further promoting research collaboration and disciplinary development. (7)(8)

In this research, we aim to develop a practical tool to build research collaboration networks using social network analysis (SNA). The tool will cover a wide range of literature data sources, store collaborations between researchers in a database, and provide search functions for different organizations, individuals, and publication networks in order to adapt to the different needs of the user's research.

Unlike traditional theories, Social Network Analysis (SNA) combines multidisciplinary convergent theories and methods from Informatics, Sociology, and Management to observe, analyze, and predict human social relationships. Based on graph theory, social network

analysis examines the structure of relationships between individuals by observing social behavior and can be applied to a wide range of fields, including mental models, market economies, transportation networks, and so on. SNA is a highly effective tool that has helped greatly sociological and psychological research (9). The two most widely used online social networking models are those that expand the network of relationships with one centrally important person at its center and perform individual centrality measurements, and those that use closed datasets for entire network construction (10). The analysis of research collaboration networks is therefore more complex. We analyze the researcher-to-collaborative network in two dimensions, starting with an individual and extending the collaborative relationship to form the network. The second is to obtain data with a limited number of links with a qualification, to construct a research collaboration network based on a topic or a subject area.

By applying the SNA technique to a comprehensive dataset of Dutch scholarly publications, the primary goal of this study is to collect accurate and reliable data about researchers' collaboration networks and to construct tools for generating collaboration networks, identifying key players and communities within them, and exploring the factors that influence the formation of the networks and their evolution over time, thus contributing to the growing number of research collaboration networks and to the dynamics of scientific collaboration in the Netherlands.

The potential areas of application of this research are vast. By mapping the co-authorship networks of Dutch researchers, this thesis will provide an intuitive way to share information about the collaboration practices and research priorities of different academic communities. In addition, the findings of this study may be useful for funding agencies that seek to support research collaborations and promote scientific innovation. By identifying key players and interdisciplinary collaboration, this study may inform resource allocation and policy development to support research networks and collaborations. In summary, this thesis will make a valuable contribution to research collaboration and will be of practical assistance to researchers, managers, and funding agencies alike.

1.1 Research Questions and Objectives

The objective of this research is to examine two broad aspects of research collaboration networks: their creation and utilization, and the evaluation of their impact, which can be summarized in the following key research questions:

- **Research Question 1: How are research collaboration networks created and utilized to improve research outcomes and inform future collaborations?**

  This question contains several sub-questions:

  - 1.1. What are the criteria and steps needed to identify appropriate research subjects and construct research collaboration networks?

  Research collaboration networks are broadly based on collaborative information from various publications, including papers, projects, software, etc., or can be defined as co-authorship. In addition to this, technical support, sponsorship, and citation relationships between researchers are often used to construct research collaboration networks. For this project, the right research subjects mean whether the final network constructed is sufficiently informative and accurate. This question focuses on the practical aspects of building a research collaboration network, such as how to obtain and process data on researcher collaboration, how to visualize the network, and how to analyze the resulting network graph.

  - 1.2. What are the key features and characteristics of these networks that can provide useful information?

  This question explores the potential uses and benefits of the collaboration network, such as identifying key researchers and institutions, visualizing patterns of collaboration, and identifying interdisciplinary collaborations. It explores what key features and characteristics of the network are most relevant to the research objectives and how these features can be visualized and analyzed to provide insights. Since nodes and relationships in social networks have many different properties, they are demonstrated in the network by changing the size, color, interaction forces, and other features of nodes and edges (11). Therefore, in this project, it is an important issue to set up the research collaboration network in a way that can provide users with clear and comprehensive access to information.

- **Research Question 2: What are the possible advantages, limitations, and biases of using network analysis to construct and evaluate collaborative research networks?**

  Sub-questions under this main question include:

  - 2.1. What are the possible biases and limitations of network analysis methods used to create collaborative networks?

This question examines the limitations and potential biases of the network analysis approach used to build the collaboration network. For example, certain types of collaborations or relationships may be underrepresented or overrepresented in the network graph, and some data sources may be incomplete or inaccurate. Understanding these limitations is important for interpreting the results of the network analysis.

- 2.2. How can research collaboration networks improve research outcomes?

This question explores the potential benefits of building research collaboration networks, such as by promoting interdisciplinary approaches, enabling access to new resources and expertise, and promoting innovation. It also explores how collaboration networks can be used to enhance research outcomes, such as by enabling access to new resources and expertise.

- 2.3. What strategies can be implemented to assess and ensure the long-term sustainability and impact of these networks?

This question aims to study how the impact and effectiveness of the collaboration network can be evaluated over time, and how it can be sustained in the long term. It also explores how the network can be updated or replicated by other researchers to ensure its validity and longevity.

1.2 Contributions

This project makes several noteworthy contributions to the field of network analysis and data visualization. First, it presents an innovative system that leverages data from multiple sources, such as Scopus, OpenAlex, and RSD, while overcoming the challenge of integrating different unique identifiers, demonstrating how different data sources can be effectively merged and exploited. In addition, the combination of Neo4j and D3 techniques in visualization provides a model for other researchers interested in the effective representation of complex networks. An example of the final view is shown in Figure 1.1.

For those who wish to replicate the study, or make improvements, the source code for the project, as well as further documentation, can be found in the GitHub repository at `https://github.com/NLeSC/rcn_py`.

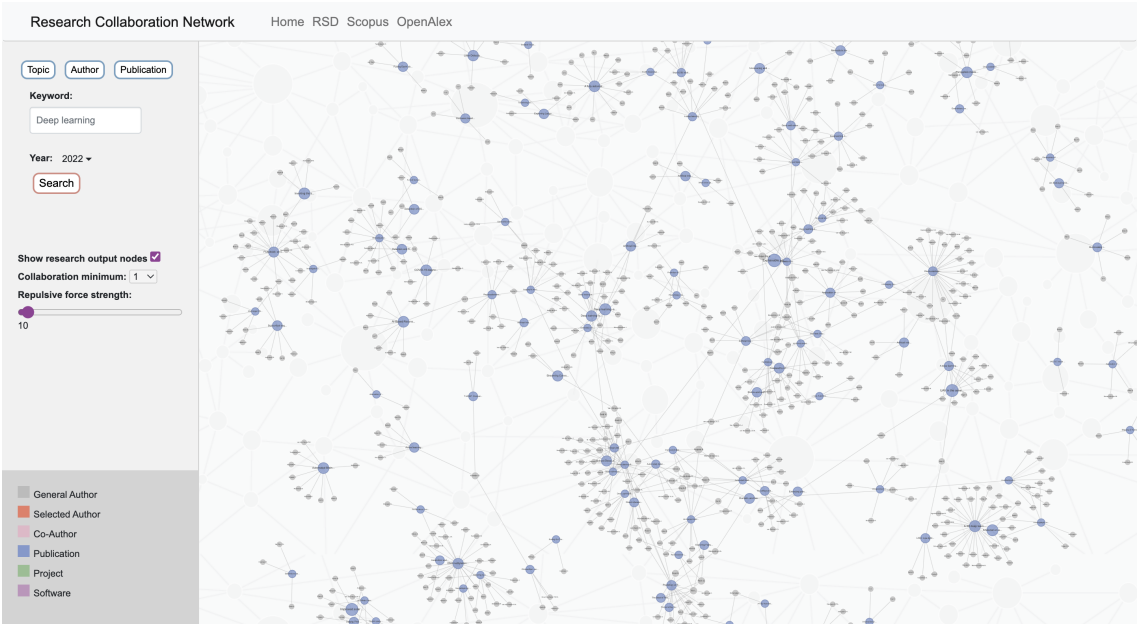


Figure 1.1: Example of the final visualization.

1.3 Outline of the Structure

This thesis is divided into nine chapters, the rest of which are developed as follows: Chapter 2 describes the background, focusing on the relevant areas: social network analysis, research collaboration networks, and author-topic modeling. Chapter 3 clarifies how our work compares to the existing literature, highlighting the unique features of our approach. Chapter 4 describes the design of our system, explaining the steps of data acquisition, processing, storage, and visualization. Chapter 5 describes our evaluation methodology, discussing how we evaluated the performance, user experience, and potential impact of our system. Chapter 6 describes our discussion of the results based on the initial research questions. Chapter 7 examines potential threats to the validity of our study. Finally, Chapter 8 summarizes our work and suggests potential future research directions.

2

Background

2.1 Social Network Analysis

Social network analysis (SNA) is a methodological strategy that has gained significant recognition in recent times. It has shown great potential for studying social interactions and networks in a variety of contexts, including the field of information science. The roots of SNA can be traced back to the mid-20th century, when social theorists began to acknowledge the influence of social structure and how patterns of interconnectedness among individuals affect social behavior. However, it was not until the 1970s that formal methodologies for social network analysis, such as graph theory and matrix algebra, began to take shape.

The late 1990s and early 2000s saw a resurgence of interest in SNA, marked by significant contributions from researchers such as Linton Freeman (12)(13), Stanley Wasserman (14)(15), and Steve Borgatti (16). In particular, Wasserman's work in developing statistical models and methods for analyzing social networks played a key role in positioning SNA as a powerful quantitative research method.

As Wasserman and Faust (1994) argued, SNA provides a set of tools and techniques for describing and understanding the structure and dynamics of social networks, such as the patterns of ties among actors, the roles and positions of individual actors in the network, and the overall properties of the network as a whole. SNA is based on the concept of a social network, which is a network of real-life individuals with social properties, that can also be understood as a knowledge graph with social properties composed of human nodes. (14)

Borgatti and Halgin (2002) (10) discussed the potential of SNA for information science research. They argue that SNA can analyze the social and structural aspects of information

processes and systems, which can help researchers to identify key actors and communities, to trace the flow of information and knowledge, and to detect barriers and opportunities for collaboration and innovation. For example, in the context of scientific research, co-authorship networks can be analyzed using SNA to identify clusters of researchers who collaborate frequently, identify the most influential researchers, and track the flow of ideas and knowledge across different research areas. This type of analysis can help researchers to identify potential collaborators, to track the evolution of research topics over time, and to understand the factors that influence scientific productivity and impact. Borgatti and Halgin also emphasize the importance of using SNA in conjunction with other research methods and techniques, such as surveys, interviews, and content analysis. By combining different sources of data and using multiple methods of analysis, researchers can gain a more comprehensive understanding of the social dynamics and processes under study.

In 2005, Carrington, Scott, and Wasserman focused on developing statistical models and methods for analyzing social networks, which helped to establish SNA as a rigorous and quantitative research methodology. A range of statistical models for social networks is discussed in the paper, including block models, p* models, and exponential random graph models (ERGMs). They also describe several computational methods for network analysis, such as centrality measures, clustering algorithms, and community detection methods. Scott's work on modeling methods in SNA helped to establish SNA as a rigorous and quantitative research methodology, and his contributions continue to be influential in the field today. In particular, his introduction of ERGMs has been particularly important, as these models allow researchers to test a wide range of hypotheses about the structure of social networks and to account for the complex dependencies that can exist in social network data. (17)

In addition to scientific research, SNA can also be applied to a wide range of other information systems, such as social media platforms, online communities, and knowledge management systems. By analyzing the social structure of these systems, SNA can help researchers to understand the factors that influence information sharing, collaboration, and innovation, and to design more effective information systems and processes.

The basic concept of SNA is to understand the relationships between the actors in a network and how these relationships affect the overall structure and dynamics of the network. Actors can be individuals, groups, organizations, or even nations. The relationships between these actors can be different types, such as social interactions, information exchange, or resource sharing. By mapping these relationships and analyzing the network

structure, SNA can provide insights into how social systems work, how information flows through networks, and how influence and power are distributed within networks (18).

One important aspect of SNA is the concept of centrality. Centrality is a measure of a node's importance within a network, based on the number and strength of their connections to other nodes. There are different measures of centrality, such as degree centrality, which counts the number of connections a node has, or betweenness centrality, which measures how often a node is on the shortest path between two other nodes in the network (19). Another important concept in SNA is clustering, which refers to the tendency for nodes to form subgroups or clusters within a larger network. Clustering can help identify important nodes or subgroups within the network.

2.1.1 Social Network Analysis Centrality

Centrality is a crucial concept in social network analysis as it measures the importance or influence of a node within a network. The centrality metrics enable us to understand how the nodes in the network are interconnected and which nodes are more important or influential than others. There are different types of centrality measures, including Degree Centrality, Closeness Centrality, and Betweenness Centrality. (20)

For node activity in a network, Degree Centrality is the simplest centrality measure, which is based on the number of direct connections a node has with other nodes in the network. Nodes with a high degree of centrality are considered to be more active or important in the network, as they have more direct connections with other nodes. However, the reality is more complex, degree centrality alone does not capture the importance of nodes that connect different parts of the network, and it is important to know whether the node is connected to people around it who are already connected or whether it is exposed to new groups (Abbasi and Altmann, 2011). (21)

Closeness Centrality, on the other hand, represents how close a node is to other nodes, and whether the direct and indirect connections between nodes provide the shortest path that allows quick access to other nodes in the network. Nodes with high closeness centrality are those that are close to other nodes in the network and have a direct or indirect connection with most other nodes.

Betweenness Centrality is another way of measuring the influence of social nodes, which is also the most commonly used centrality measure. Betweenness centrality takes into account the node's role as a connector in the network, where a node resides in the network when it acts as a mandatory path for two other nodes to connect (Leydesdorff, 2007). (22) It measures the number of times a node lies on the shortest path between two other

nodes. Nodes with high betweenness centrality are critical in maintaining the network's connectivity, and their removal can disrupt the flow of information in the network (Salter-Townshend, 2012). (23)

Centrality measures in SNA provide valuable insights into how nodes are interconnected in a network and their relative importance or influence. The intuitive visualization of different centrality metrics is shown in Figure 2.1, a notably red node signifies a high centrality score. By understanding the centrality of nodes, we can identify key players, influential actors, and potential bottlenecks in the network, which can help in decision-making, resource allocation, and network interventions.

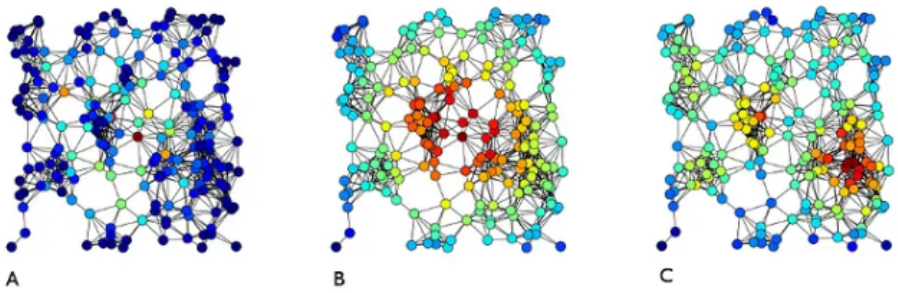


Figure 2.1: Comparison of A) Betweenness, B) Closeness, C) Degree centrality based on Wikipedia.

2.1.2 SNA Clustering

SNA clustering is a method used to identify clusters or sub-groups within a social network. Clusters are groups of nodes that are more densely connected than nodes in other parts of the network. In other words, clusters are groups of nodes that tend to interact more frequently with each other than with nodes outside the cluster.

In order to understand collaborators and their social networks, we need to evaluate node locations and cluster distributions, and understand information about the leaders, connectors, isolates, and groups in the network. One way to do this is to evaluate the centrality of participants in the network.

Clustering refers to the tendency of individuals in a network to form groups or clusters based on shared characteristics or interests. Clustering can reveal hidden structures and sub-communities within a network and help understand the dynamics of information and resource flow within and between clusters. There are different types of clustering algorithms, such as k-means (24), hierarchical clustering, and spectral clustering, that can be

used to identify clusters in social networks. These algorithms can help researchers identify patterns and relationships within the network that may not be immediately apparent. (25)

2.2 Research Collaboration Network

Research collaboration networks are a complex and dynamic system of relationships between individuals or organizations involved in research activities. The concept of research collaboration networks has been widely studied in various fields, including sociology, information science, and bibliometrics. These networks can be described as a social structure that is formed when researchers collaborate to work on a project or paper. The study of these networks can provide valuable insights into the dynamics of scientific collaboration and the diffusion of knowledge.

The history of research collaboration networks dates back to the early 20th century when bibliometrics emerged as a field of study. Bibliometrics is the quantitative analysis of bibliographic data, including citation analysis, co-authorship analysis, and publication analysis. In the 1960s, sociologists began to study the social networks of scientists, which led to the development of the field of social network analysis (SNA). The growth of research collaboration networks is driven by a variety of factors, including the increasing complexity of research projects, the need for interdisciplinary collaboration, and the globalization of science. (26)

Several types of research collaboration networks capture different dimensions of scientific collaboration as follows:

- **Co-authorship Networks:** Co-authorship networks are formed based on collaborations between researchers who have jointly authored scientific publications. These networks represent the connections between researchers who have collaborated on research projects, indicating the strength and frequency of their collaborations. Figure 2.2 presents an example of Newman's coauthorship networks (2004). Analyzing co-authorship networks can reveal key actors, influential research groups, and communities of researchers within a specific field or domain. (7)

- **Citation Networks:** Citation networks are constructed based on the citations among scholarly publications. These networks illustrate the relationships between publications through citation links, where one publication cites another (de Solla Price, 1965) (27). By analyzing citation networks, researchers can identify influential

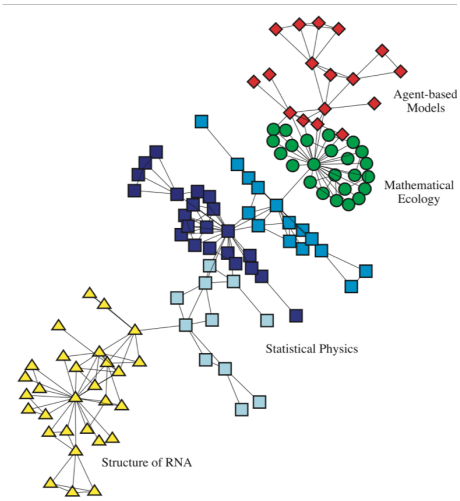


Figure 2.2: An example of a small coauthorship network depicting collaborations among scientists at a private research institution. Nodes in the network represent scientists, and a line between the two of them indicates they coauthored a paper during the period of study. (1)

publications, renowned authors, and research communities with high citation impact, which provide a view of the connections of research in the field.

- **Co-citation Networks:** These networks represent collaborations based on the co-citation patterns of publications. Researchers who are frequently cited together in the literature are considered to have collaborative relationships (Small, 1973) (28). Figure 2.3 shows two of the co-citation networks of Alnajem *et al.*. Analyzing co-citation networks can reveal clusters of closely related research and identify influential researchers or research groups.

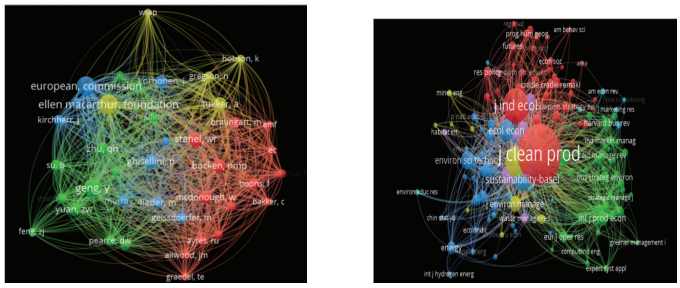


Figure 2.3: Co-citation by author/source network in circular economy research. (2)

- **Patent Collaboration Networks:** In fields related to innovation and technology,

collaboration networks based on patent co-inventorship can be constructed. These networks capture collaborations between inventors who jointly file patents, providing insights into collaborative innovation and technological advancements. (29)(30)

- **Collaboration Networks based on Funding:** These networks represent collaborations between researchers or research groups who have received funding from the same sources or participated in joint research projects. By analyzing such networks, researchers can identify patterns of funding collaboration and explore the impact of funding agencies on research collaborations. (26)

- **Collaboration Networks based on Affiliation:** These networks focus on collaborations between researchers affiliated with specific institutions, universities, or organizations. They provide insights into the collaborative relationships within and across different academic or industrial entities and shed light on the collaborative patterns among researchers associated with specific affiliations. (31)(32)

- **Collaboration Networks based on Geographic Proximity:** Sonnenwald (2007) discusses the impact of geographic factors on scientific collaboration, including how they contribute to the formation of regional research clusters. (33) These networks capture collaborations between researchers or institutions based on their geographic proximity. By examining such networks, researchers can understand the dynamics of local collaborations, regional research clusters, and the influence of geographic factors on collaborative relationships.

2.3 Collaboration Networks Topic Modelling

The size of the network must be limited for the analysis of researcher social networks, and we typically limit the range of network nodes to the same topic, i.e. researchers who study similar topics. Most search engines can provide search by keyword, but many papers have vague keyword extraction, or keywords are extracted by machine with errors. This is why it is useful to find research topics by modeling the topic of the abstract or body of the paper, or to explore the relevance of the author to the topic of the literature by modeling the author's topic of the researcher. For topic clustering models of research collaboration networks, some relevant topic modeling techniques are outlined in the Literature Study in this project. In this section, the underlying logic of the technique is briefly described.

For traditional document topic classification, a popular and widely used approach is Latent Dirichlet Allocation (LDA) (Blei et al., 2003) (34). LDA is a generative probabilistic

model that allows the identification of potential topics in a corpus of documents. The traditional LDA model assumes that documents are composed of a mixture of topics, each represented by a distribution of a set of words. The underlying assumption is that documents are generated by selecting topics and then selecting words from those topics. LDA uses Bayesian inference to estimate the topic-word distribution and the proportion of topics in each document.

While traditional LDA models are powerful for identifying topics within a collection of documents, they do not explicitly capture the influence of authors on the content of the documents. In many cases, the author's expertise, interests, and writing style play a significant role in shaping the topics present in their documents. Author-Topic Models (ATM) extend LDA by incorporating authorship information into the topic modeling process. (35)(36)

Rosen-Zvi et al. (2012) (3) proposed an LDA-based model for document collection generation, the Author-Topic Model. This can be done by modeling information such as the content of the document and the author's research area to obtain the topic set of the corpus as well as to identify the topics used by the author. The LDA model consists of two sets of unknown parameters, the document distribution, and the subject distribution. As shown in Figure 2.4. Modeling information such as the content of the document and the author's research area, is combined to obtain the topic set of the corpus and to identify the topics used by the author.

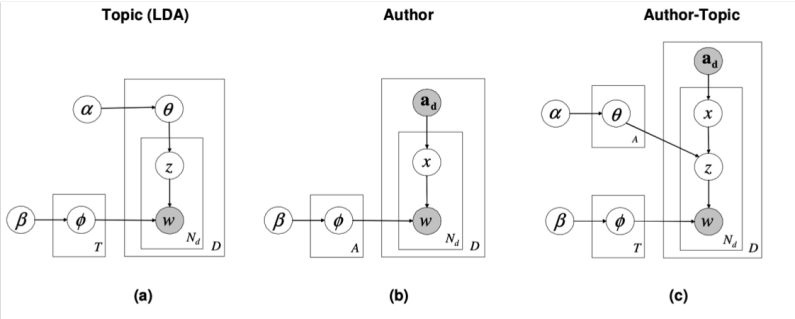


Figure 2.4: Comparison of LDA topic model, author model, and author-topic model. (3)

By utilizing author-topic models, researchers and practitioners can gain deeper insights into the interplay between authors, topics, and documents. If comprehensive author information is available in the dataset, the ATM could provide a richer understanding of the underlying structure of textual data, facilitating more accurate document categorization, topic labeling, and personalized recommendations.

3

Design

The design section of this thesis provides a detailed description of the system or solution developed for analyzing co-authorship networks and conducting topic modeling using Scopus, OpenAlex, and RSD (Research Software Directory) data. This section focuses on explaining the design choices made throughout the development process, outlining the architecture, data flow, and key components of the system. The design ensures the effective integration of D3 visualization, Neo4j database, and topic modeling techniques to achieve the research objectives.

Figure 3.1 provides a high-level outline of the proposed co-authoring network application system. It delves into the design and implementation of the system architecture and interactions between different solution components in the process. The system follows a modular architecture that integrates various components to enable efficient co-authorship network analysis. The main components include data extraction and preprocessing, network construction, database integration, visualization, and topic modeling. These components work together to provide a seamless and user-friendly experience.

The data extraction and preprocessing component acquire data from bibliography resources. It utilizes appropriate APIs, or data access mechanisms to retrieve relevant information. The extracted data then undergoes preprocessing tasks such as cleaning, deduplication, and standardization to ensure data quality and compatibility. The database facilitates the efficient storage, querying, and retrieval of co-authorship network data. It integrates with a Neo4j graph database, which provides a powerful platform for managing graph-based data and will be discussed in Section 3.3.

The network construction component takes the preprocessed data as input and generates co-authorship networks. It identifies authors and their co-authorship relationships, constructing a network graph representation. Various algorithms and techniques are applied

to capture meaningful collaboration patterns and relationships within the network.

The visualization utilizes the D3 JavaScript library to create interactive and visually appealing visualizations of co-authorship networks. It takes the network data stored in the database and dynamically renders the network graph. Users can explore the network, visualize collaboration patterns, and interact with the visualizations to access detailed information about authors and publications.

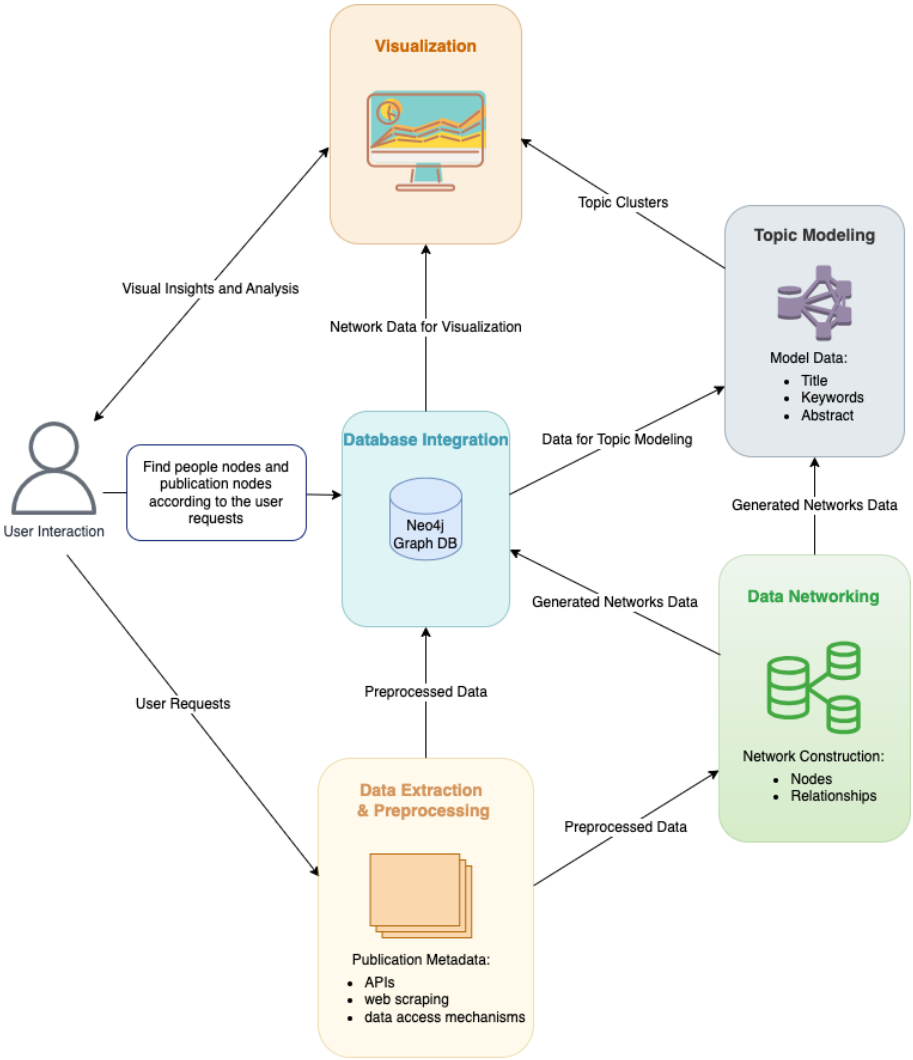


Figure 3.1: Co-authorship Networks System Architecture Diagram

3.1 Data Collection

When selecting the publication data sources for the co-authorship network analysis, several options were considered to ensure comprehensive and reliable metadata. The chosen sources for obtaining publication data include Scopus, Google Scholar, CrossRef, and OpenAlex. Each source has its advantages and limitations in terms of data accessibility and coverage. In addition to the above datasets, we explored datasets such as Springer, IEEE, DBLP, and Semantic Scholar, but abandoned them at a very early stage as they all lacked the necessary metadata we needed to build our collaborative network, e.g., Springer and Semantic Scholar lacked information on nationalities and affiliations, and IEEE and DBLP were specific to only a single subject area.

- **Elsevier Scopus:**

  Scopus was identified as a primary data source due to its extensive coverage of scholarly publications across various disciplines. It provides a wealth of metadata including author affiliations, publication details, and citation information. (37) However, accessing Scopus data usually requires subscribed access, limiting its availability to users outside academic institutions. Additionally, while Scopus offers an API package called 'pybliometrics', it has strict request limits (20,000 per week), allowing only a limited number of requests per week. (38) This can be a constraint when dealing with a large volume of data. To overcome this limitation, manual downloading of metadata files from Scopus is recommended, ensuring the collection of a comprehensive set of publications.

- **Google Scholar:**

  Google Scholar is widely used as a source of scholarly publications. While it does not have an official API, there are several Python packages available, such as 'scholarly', that enable the scraping of Google Scholar data. However, it should be noted that Google Scholar's data may not provide all the necessary information required for building a complete co-authorship network. Essential properties such as author affiliations and citation details may be limited or not readily available. Therefore, Google Scholar may serve as a supplementary source of publication metadata, complementing the data obtained from other sources.

- **CrossRef:**

CrossRef is a valuable resource for obtaining publication metadata, especially due to its association with ORCID (Open Researcher and Contributor ID). ORCID provides a unique identifier for researchers, allowing for more accurate author disambiguation. CrossRef's integration with ORCID ensures that author information is based on a widely-used and standardized identifier. However, building a complete co-authorship network using CrossRef data can be time-consuming. This is primarily due to the necessity of making a large number of requests to retrieve all the relevant publications. Additionally, CrossRef does not offer a comprehensive static dataset, limiting the ability to create a complete dataset solely from CrossRef data.

- **OpenAlex:**

  OpenAlex offers a distinct feature for searching works affiliated with a specific institution, making it particularly useful for analyzing co-authorship networks based on affiliations. While OpenAlex provides extensive author data, including author names, affiliations, and publication records, it has limitations in terms of the availability of ORCID values. (39) Many ORCID values in OpenAlex are null, which can impact the completeness and accuracy of the co-authorship network. As a result, OpenAlex is often used as an additional source of information for affiliation-based analyses, supplementing data obtained from other sources.

- **Research Software Directory:**

  Research Software Directory (RSD) is a comprehensive platform for sharing and cataloging research findings, containing software and projects from various institutions. It is a rich repository for research outputs from the Netherlands e-Science Centre.

By integrating data from Scopus, Google Scholar, CrossRef, OpenAlex and RSD (a brief comparison and description is shown in Table 3.1), a more comprehensive dataset can be constructed for the co-authorship network analysis. This combination ensures a broader coverage of publications and collaboration patterns among authors from different sources. Careful consideration of the strengths and limitations of each source allows researchers to leverage the available data effectively, gaining valuable insights into co-authorship relationships within the scholarly domain.

Dataset	Advantages	Disadvantages	Used in Final Application
Elsevier Scopus	Extensive coverage of scholarly publications; offers rich metadata	Requires subscribed access; has strict request limits	Yes
Google Scholar	Widely used; numerous Python packages available for data scraping	Does not provide comprehensive metadata; no official API	No
CrossRef	Integration with ORCID ensures standardized author information	Time-consuming to build a complete network; no comprehensive static dataset	No
OpenAlex	Supports affiliation-based searches; offers extensive author data	Many ORCID values are null	Yes
RSD	Comprehensive platform for sharing research outputs; focus on Dutch institutions	Limited to software and projects; lacks other research records	Yes

Table 3.1: Summary of Datasets Used in the Study.

3.2 Data preprocessing and cleaning

3.2.1 Removal of Duplicate Publications

To remove redundancy and ensure accuracy, the dataset undergoes a process of identifying and removing duplicate publications. A duplicate detection algorithm is used to compare various attributes such as DOI (Digital Object Identifier), and, if no DOI is available, attributes such as title, author, and date of publication are used for comparison.

Removing duplicate publications is crucial to streamline the dataset and prevent later bias. Duplicate entries can miscalculate collaboration counts and generate inaccurate information in the co-authorship network analysis. By removing duplicate entries, the accuracy and reliability of the co-author network analysis are improved, enabling a more accurate understanding of the patterns of collaboration and relationships between authors.

3.2.2 Standardization of Author Names

In order to address the problem of different representations of author names in different publications, we need to standardize the authorship of names. This process ensures that authorship is consistent. String matching algorithms, are used to identify authors with similar or identical names. we need to compare attributes such as first name, surname, initials, and even affiliation to identify potential matches. This merging of authorship

reduces redundancy and ambiguity in the dataset, allowing for a more accurate analysis of the co-authorship relationship.

3.2.3 Disambiguation of Author IDs

In data collection and analysis, it is vital to distinguish between researchers. Due to duplicate names, a unique ID such as ORCID needs to be identified. however, the problem is that while ORCID is the most recognized researcher code, only CrossRef has the majority of ORCID records. Scopus author-specific IDs are defined as scopus_id, and only a minority of researchers fill in their own ORCID, making it challenging to use the API to find the correspondence between scopus_id and ORCID becomes a challenge.

Is it possible to avoid this problem by examining only Scopus data and using scopus_id as a unique identifier? Unfortunately, our study looks at both publication data and Research Software Directory (RSD) data, where author IDs are defined using ORCIDs. Therefore, we need to use multiple search methods to merge the same authors in both datasets.

First, the RSD dataset is much smaller than the Scopus dataset, so retrieving the scopus_id from the ORCID of the RSD is less computationally intensive. It is worth noting that the researcher data in RSD consists of three attributes: name, ORCID, and affiliation, and that both ORCID and affiliation may have null values. Although the Scopus Author API provides search functionality for these three attributes, it is important to consider that many people may have different names registered on RSD and Scopus, with variations such as middle name or preferred name. In addition, there may be cases where affiliations are not updated, resulting in inconsistent information on Scopus and RSD, which is also addressed in the "Threats to validity" chapter of this thesis 7.

3.2.4 Filtering of Non-Relevant Publications

To ensure that the co-authorship network analysis accurately reflects the desired research context, it is essential to address non-relevant publications and consider specific criteria for inclusion in the dataset.

Non-relevant publications, such as book chapters, editorials, or conference proceedings, are outside our target domain. Therefore, a filtering process is implemented to exclude these non-relevant publications from the dataset. For topic modeling, it is important to focus on publications that are primarily in English, which ensures linguistic consistency across the dataset and accurate topic extraction and clustering. In addition, given the

nature of collaborative networks, particular attention is paid to publications with more than two authors, as co-authorship networks primarily capture collaborative relationships.

3.3 Graph Database Storage

In co-authorship network analysis, efficient data processing relies heavily on the storage and management of graph datasets. These datasets reflect the complex relationships and connections between authors, publications, and affiliations. Therefore, choosing a suitable storage method is crucial to ensure smooth data retrieval, query, and analysis.

There are several storage options for graph datasets, including relational databases, document-oriented databases, and graph databases. Each has its advantages and is suitable for different requirements. Relational databases are suitable for handling complex queries, and document-oriented databases are suitable for handling textual data (such as additional metadata related to publications or authors). (40)

Among these databases, graph databases like Neo4j (41) are widely adopted due to their special design for storing and managing graph structures. They provide efficient traversal and query capabilities to accommodate the dynamic nature of the Web. (42) With Neo4j, co-authorship networks can be efficiently represented as graphs with nodes and relations, while Cypher, a powerful query language, can retrieve specific subgraphs and explore collaborative networks.

3.3.1 Data Parsing and Networking

In the context of co-authorship network analysis, the process of data parsing and networking plays a crucial role in organizing and establishing the relationships between authors and publications. This subsection presents a detailed description of how author information, publication information, and their relationship "IS-AUTHOR-OF" are saved using the acquired metadata. Notably, the explicit storage of "coauthor" relationships between authors is deemed unnecessary due to the potentially high number of links, while the "IS-AUTHOR-OF" relationship adequately captures co-authorship information and can be leveraged using Cypher queries.

- **Saving Author Information:** To preserve the author-related data, including names, affiliations, and ORCID identifiers, an author node is created for each unique author within the dataset. The author node serves as a repository for attributes such as the

author's name, affiliation, and ORCID identifier, enabling a comprehensive understanding of the author's background.

- **Storing Publication Information:** Similarly, the metadata obtained contains important publication-related data such as title, publication date, and DOI. Each publication record generates a publication node that stores the basic attributes of the publication such as title, publication date, and DOI.
- **Establishing the "IS-AUTHOR-OF" Relationship:** The basic link between authors and publications is achieved by establishing an "IS-AUTHOR-OF" relationship. Within the framework of a graph database, this relationship is represented as a directed edge connecting an author node to the corresponding publication node.

Neo4j ensures that essential author information, publication details, and the associations between them are stored appropriately. The decision not to store 'co-author' relationships between authors is based on network complexity considerations. Instead, the 'IS-AUTHOR-OF' relationship provides an efficient means of capturing co-author information, allowing efficient use of Cypher queries (as shown below).

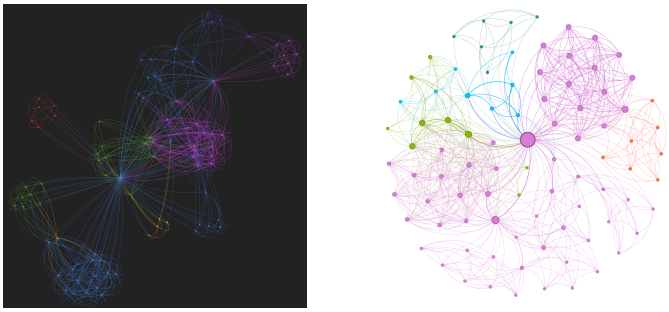
```
MATCH (m:Person)-[:IS_AUTHOR_OF]->(p:Publication)<-[:IS_AUTHOR_OF]
      -(n:Person)
RETURN m, n
```

3.4 Visualization Design

The visualization design plays a pivotal role in presenting the co-authorship network analysis results in a clear, intuitive, and visually appealing manner. After exploring several visualization tools, including Pyvis (Figure 3.2(a)), Gephi (43) (Figure 3.2(b)), and Cytoscape (44), the decision has been made to leverage the power and flexibility of D3.js for the visualization of the co-authorship network, with the final visualization shown in Figure 3.3.

By utilizing HTML, CSS, and SVG, D3.js provides a flexible and customizable framework for describing complex network structures and relationships (45). Visual interactivity allows users to gain greater insight into the patterns of collaboration and connections between authors.

For the representation of nodes, authors, and publications are the main entities in a co-authorship network, and we use nodes to represent these entities. In the visual design, authors and publications are differentiated by color, as shown in Figure 3.4. The size and



(a) Pyvis (b) Gephi

Figure 3.2: Examples of visualization tools Pyvis and Gephi.

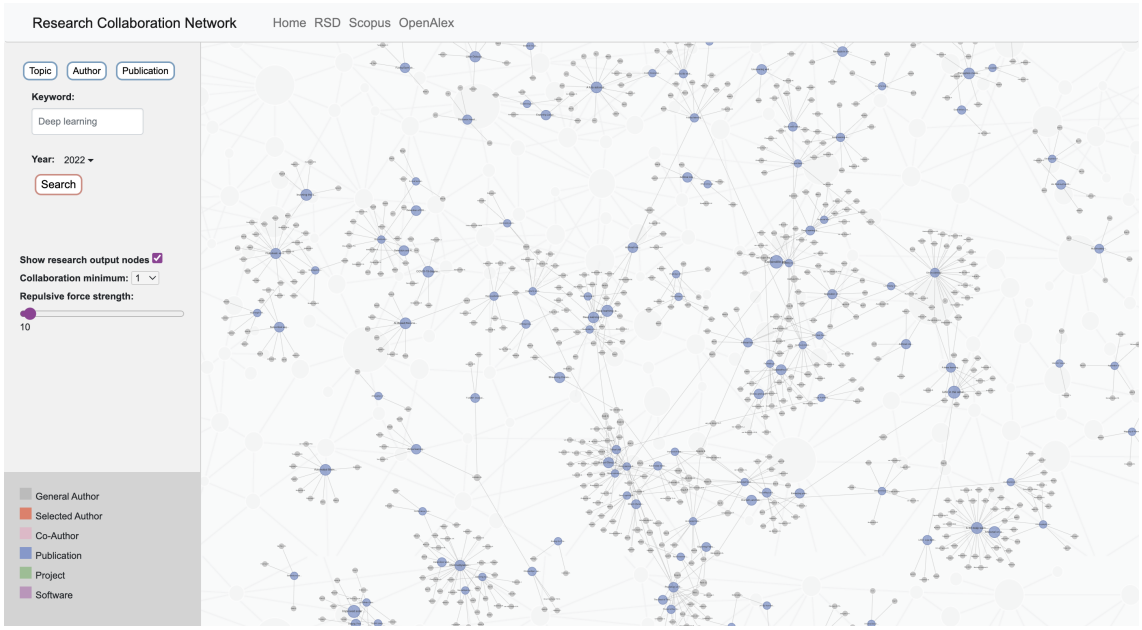


Figure 3.3: Example of the final visualization using D3.

color of the nodes can be used to represent additional information such as the number of times the researcher appears in the network, the number of times the publication has been cited, etc. The relationships between authors and publications are visualized by the "IS-AUTHOR-OF" relationship, using edges or links. These links illustrate the co-authorship relationship. The edges can be customized in terms of thickness, color, and transparency to indicate the strength of the different relationships.

Determining an effective layout and arrangement of nodes is important for aesthetic visualization. d3.js provides various layout algorithms, such as force-directed layout, which

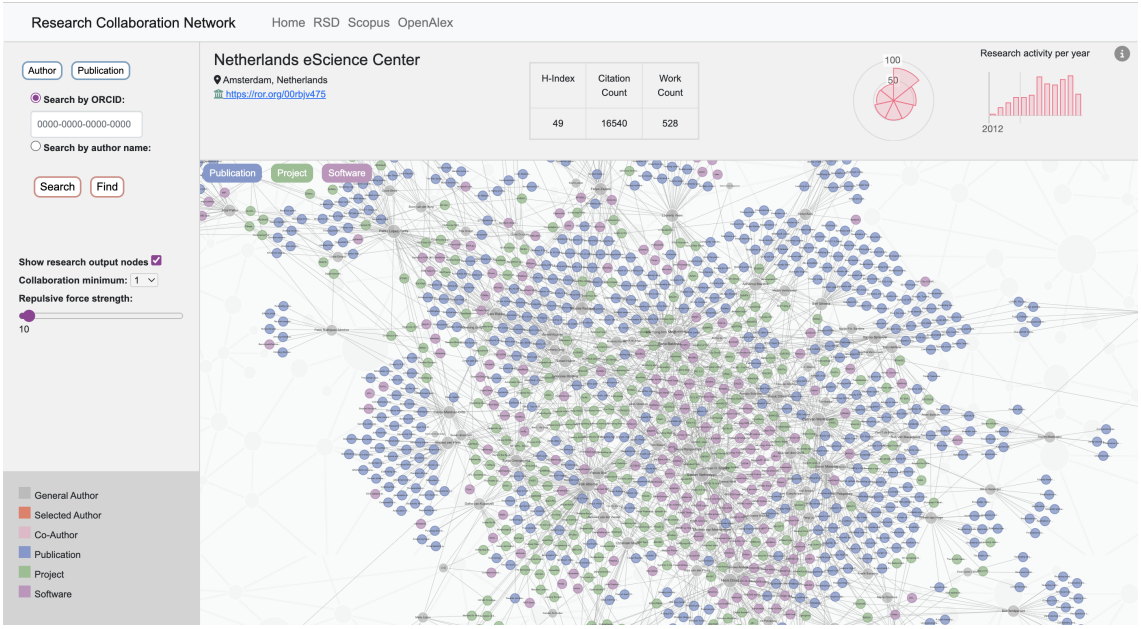


Figure 3.4: Examples of different nodes.

simulates physical forces and positions nodes according to the forces of attraction and repulsion between them. This approach ensures that highly interconnected authors and publications are visually grouped, facilitating the identification of clusters and communities within a network of co-authors. In addition, zooming and panning functions are integrated into the visualization to accommodate networks of different sizes. Users can zoom in and out of the network to examine specific areas of interest in detail or to get a broader overview. The panning function in the visualization allows the exploration of large-scale networks without sacrificing the contextual understanding of the entire co-authored network.

Interactivity is a key aspect of visualization design, enhancing the user experience and enabling users to explore more information. Inspired by the Neo4j browser (46), tooltips can be implemented to provide contextual details about authors and publications, as shown in Figure 3.5. When the mouse hovers over a node, the tooltip can display information such as the author's name, affiliation, or the title and publication date of the publication, as well as provide the ability to hide the node, fix the position of the node and show all relationships for that node. This interactive functionality allows the user to access specific details without cluttering the visualization.

By adopting D3.js as the visualization framework, we have designed the results of the co-authorship network to be visually effective in communicating with the user. This visualization has been designed to enable users to gain insight into patterns of collaboration,

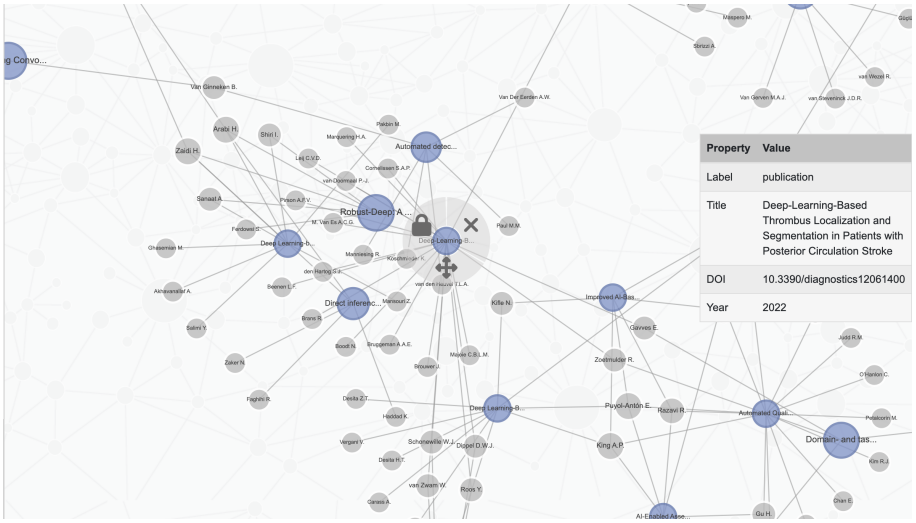


Figure 3.5: Example of Tooltips in the D3 network.

identify influential authors, explore clusters, and examine the co-authorship network in greater depth.

3.4.1 Web Integration for Data Retrieval

The interaction between HTML, JavaScript, and Python constitutes a fundamental aspect of the system that facilitates user queries, data retrieval, and database processing.

At the user level, HTML is used to capture user queries and present web visualizations. JavaScript acts as an intermediary between the user's input and server-side functionality. By adding an event listener to the HTML form, JavaScript handles the events triggered when the user submits the form. The default form submission behavior is then circumvented and the subsequent steps of the interaction are initiated.

After the form is submitted, JavaScript instantiates a request provided by the XMLHttpRequest object (47). This request is directed to a specific route in the Python Flask server to accommodate the search function. After receiving the query parameters from the request object, the Flask function takes on the basic task of query processing. The Flask function then establishes a connection to the database, executes the query, applies filters, and processes the retrieved data into JSON format to facilitate seam data exchange between the Python Flask server and the JavaScript code. In the JavaScript code, the XMLHttpRequest object's "onreadystatechange" event handler focuses on monitoring changes in the request state. Upon receiving a response from the Flask server, the handler processes the response data.

Using the processed response data, JavaScript dynamically updates the D3 network and text information on the user interface, enabling the presentation of search results or visualizations.

3.4.2 Scopus Publication Data Network

The system includes three different types of networks, 'subject networks', 'author networks', and 'publication networks', each of which meets specific user search requirements and allows for focused exploration of co-authorship. This section outlines the design choices and associated functionality of each network.

- **Topic Network Search:** Researchers often seek to investigate collaborations within a particular field or area of interest. By providing the ability to enter keywords and specify time intervals, users can narrow their focus and obtain a network that captures publications relevant to their chosen topic.

  The topic network is designed to contain no more than 300 publications. This limit provides both a comprehensive view of co-authorship relationships within the selected topic and ensures that the visual information is readable. Figure 3.3 presents the topic network related to Deep learning in 2022.

- **Author Network:** The Author Network Search allows users to conduct an author search using either the author's ORCID (Open Researcher and Contributor ID) or their full name.

  When users initiate an author search, the system handles variations or discrepancies in author name representations. Author names can differ due to factors such as name variations, alternative name spellings, or different formatting conventions across publications. By flexibly processing the author's name, the Author Network accommodates different representations of the same author. This approach mitigates the risk of missing relevant co-authorship relationships due to inconsistencies in author name formatting.

- **Publication Network:** The Publication Network allows users to explore co-authorship relationships based on specific publications. Users can conduct a publication search using the DOI (Digital Object Identifier) of a publication of interest. This search enables the retrieval of relevant co-authorship relationships associated with the selected publication, providing insights into the collaborative network surrounding the publication.

The design of these networks emphasizes usability and efficient retrieval of relevant co-authorship information. By focusing on Scopus Publication Data, which encompasses English papers published in the last ten years, the networks provide up-to-date and comprehensive insights into the co-authorship landscape. The use of specific search parameters, such as keywords, time intervals, ORCID, and DOIs, facilitates targeted exploration and analysis.

3.4.3 OpenAlex Data Retrieval

The data retrieval process from OpenAlex plays an important role in the research. Because it has an extensive and comprehensive institutional publishing dataset, OpenAlex provides a convenient way to access institutional publishing records, which we use to build our institutional network visualization.

To account for the extensive data size, we designed an interface where users can specify a time range for their search by setting a start year and an end year.

The data from OpenAlex also serves a key function in topic modeling for our visualization. Each publication in the OpenAlex dataset includes a 'concepts' classification, which we used to perform topic modeling based on their topics, with each cluster represented by a unique color in the visualization. Moreover, we adjusted the force strength between nodes within the same group to visually aggregate them into clusters, enhancing the user's ability to perceive topic-based patterns in the network.

3.4.4 NLeSC networks from RSD

Our research also makes use of the Research Software Directory (RSD), a comprehensive data source on published and unpublished software and projects from Dutch institutions. This unique repository provides a valuable record of the Dutch collaborative network and forms the basis of the Dutch e-Science Center (NLeSC) network in our study.

To merge the RSD data with the pre-existing Scopus data in our database, we needed a common identifier that spanned both datasets. For this purpose, we used the ORCID (Researcher's Unique Identifier) as a bridge. We retrieved all work from the RSD for the e-Science Center and proceeded to match authors to their respective ORCIDs. Since many of the ORCID fields in the RSD data were empty, we had to devise an alternative strategy to accurately identify authors.

To overcome this challenge, we used the author names and affiliations provided in the RSD data to query the ORCID API to retrieve the corresponding ORCIDs for each author.

Subsequently, we used these ORCIDs to query the Scopus database to check for any author records that might exist in it.

Users can select a node in our visualization to extend its connectivity. Once selected, the system retrieves and displays all relationships associated with the selected node from the Scopus and RSD datasets. This feature provides the user with a view of the research in different databases.

4

Evaluation

In this chapter, we present an evaluation of the Research Collaboration Web Service. Our evaluation focuses on aspects such as the performance of the system, the user experience it facilitates, and its potential impact on the field of academic research.

4.1 System Performance

The performance of our systems can be measured quantitatively by specific parameters. Given the large amount of data our project requires to process, response time, the speed with which the system responds to user queries, becomes an important metric for measuring performance.

We assessed this aspect through several trials, conducting different types of searches including topic, author, publication, and institution searches. Each trial was repeated three times to obtain an average response time. The results are detailed in Tables 4.1 to 4.4 below. As can be seen, the response times are still within acceptable time ranges even when dealing with large networks.

Network Size (Number of Publications)	Trial 1 (ms)	Trial 2 (ms)	Trial 3 (ms)	Average Response Time (ms)
100	694.5	621.2	760.1	691.93
200	868.8	921.2	927.0	905.67
300	1051.3	1160.3	1057.2	1089.60
400	948.0	1216.9	1182.6	1115.83
500	1042.5	1204	1175.1	1140.53

Table 4.1: Average Response Time per Network Size for Topic Searches

Table 4.5 shows the time taken to generate the D3 network diagram after acquiring the processed network data. The effect of increasing the number of publications from 50 to 500 is recorded.

Number of Nodes in Network	Trial 1 (ms)	Trial 2 (ms)	Trial 3 (ms)	Average Response Time (ms)
663	2347.1	1840.5	1730.4	1972.67
1593	2260.8	1887.4	1749.3	1965.83

Table 4.2: Average Response Time for Author Searches

Trial 1 (ms)	Trial 2 (ms)	Trial 3 (ms)	Average Response Time (ms)
1422.8	1875.9	1754.6	1684.43

Table 4.3: Average Response Time for Publication Search

Trial 1 (ms)	Trial 2 (ms)	Trial 3 (ms)	Average Response Time (ms)
2099.2	2833.3	3411.3	2781.27

Table 4.4: Average Response Time for Institution Search

The results clearly show that as the number of publications in the network increases, the average generation time of the D3 network graph also increases. This is an expected result since the complexity of the computation increases as the number of nodes and links increases. However, it is important to note that the visualization of networks with more than 500 publications becomes meaningless to the user due to the high number of nodes (more than 2,000) and links (more than 4,000). Therefore, we limit the performance evaluation to networks with 500 or fewer publications.

Number of publications	50	100	150	200	250	300	350	400	450	500
Mean Generation Time (ms)	52.7	114.0	194.3	212.3	243.0	280.0	277.7	291.0	287.7	298.3

Table 4.5: D3 Network Graph Generation Time

4.2 User Experience

A successful system is not just about the performance it delivers, but also the user experience it facilitates, although this aspect is less amenable to quantitative measurement. The effectiveness of our system's user interface, the clarity of the visual images generated, and the user-friendliness of the system's interface are all areas worth evaluating.

- **Ease of Use:** Our services are designed to be simple and easy to use. Our service was designed with simplicity and ease of use in mind. The search interface is intuitive, making it easy for users to find publications or authors. Search results are presented in visual charts that can be easily manipulated and explored, making the discovery process more engaging and informative.

- **Performance:** One of the challenges of building research collaboration networks is dealing with large networks with many nodes and links. Our service handles this problem through topic modeling and temporal filters. This allows users to manage the complexity of large networks and focus their attention on the most relevant parts of the network.

  As we discussed in the previous section, the generation time of the D3 network graph plays an important role in the user experience. From our performance analysis, we found that the generation time for networks with less than 500 publications is within acceptable limits. This ensures a responsive and smooth user experience, which contributes to overall user satisfaction.

- **Visual Appeal:** The visual design of the web graph was attractive and informative. Different colors are used to represent different topics, making it easy for users to discern the structure and relationships within the network. The ability to zoom in and out allows users to explore the details of the network at different levels.

- **Flexibility:** Our service provides flexibility in searching and exploring the Web. Users can search by topic, author, or institution and filter results by year. This flexibility allows users to customize the web to their specific interests and research needs.

While this evaluation is based on the expected user experience, future work should include a thorough user experience evaluation of actual users. Their feedback can provide valuable insights and guide further improvements to the service.

4.3 Impact and Usefulness

The ultimate test of our service is its practical impact and usefulness, which can be assessed by understanding the tangible benefits it provides to its primary users (researchers and institutions). Key aspects to be assessed here include the effectiveness of the service in identifying potential collaborators, its ability to elucidate research trends, and the value it adds to the decision-making process for future research collaborations.

While our service is designed to help users identify potential collaborators, we also hope to illuminate trends in research collaboration by visualizing the ebb and flow of collaboration networks. While it is challenging to quantify this aspect directly, user feedback and

evaluations can provide valuable insights. For example, if users report success in identifying and engaging with new collaborators through our service, this would indicate a positive outcome.

5

Discussion

In this chapter, the following discussion presents the results of our study in the context of our initial research questions. We examine the implications of our research on co-authorship network analysis, drawing from multiple data sources including Scopus, OpenAlex, and RSD data, with the aid of the Neo4j database and D3 visualization tool.

- **Research Question 1: How are research collaboration networks created and utilized to improve research outcomes and inform future collaborations?**

  Building a research collaboration network is a complex process, which is described in detail in Chapter 3.

  - 1.1. What are the criteria and steps needed to identify appropriate research subjects and construct research collaboration networks?

  To determine the appropriate research subjects for a research collaboration network, it is important to consider the purpose of the network and the type of collaboration model to be explored. As discussed in Section 2.2 2.2, there are several types of collaborative networks, including co-authorship networks, citation networks, funding networks, and affiliation networks. While all of these models provide valuable information about the dynamics of collaboration, implementing them simultaneously can be a challenge. In our study, we chose to focus on co-authorship networks, which provide rich insight into the history of collaboration by using authors and publications as nodes and relationships between them as links. We chose this approach because it illustrates direct collaboration between researchers and produces an informative and comprehensive visual representation of the research field.

In terms of criteria for topic selection, we prioritize publications with complete metadata, such as those that are published, provide comprehensive information, and have searchable author information. Starting with one publication allows for the creation of an exhaustive network, even if it starts relatively small. This approach ensures that the resulting collaborative network is not only accurate but also detailed and insightful, providing a strong basis for subsequent analysis and research.

Through our work, we have established a framework for constructing research collaboration networks. A crucial first step involves data acquisition and cleaning from diverse sources. Subsequent data integration using tools like Neo4j enables the creation of the foundational network. With D3 visualization, we could effectively visualize the intricate network structures. This combination of processed data and advanced tools provides a viable method for building research collaboration networks.

- 1.2. What are the key features and characteristics of these networks that can provide useful information?

In constructing the research collaboration network, we've developed a visualization platform that offers multi-perspective search options across different data sources. This feature-rich environment caters to users' varied needs, allowing them to extract various information based on their specific search requests.

Topic-Based Focus:

Users can explore the landscape of a particular research topic, identifying key contributors, groundbreaking work, and trends over time. This information is critical to understanding the evolution of a research field and identifying potential gaps or opportunities for future work.

Figure 5.1 shows the "deep learning" topic networks for the period 2014-2022. The topic-based approach effectively tracks the evolution of a research area, showing its growth and the emergence of new trends

While topic search is useful, it has limitations. It is title- and keyword-based, and it only pulls from the Scopus database, which may miss relevant research. Future improvements could include refining the search algorithm and incorporating more databases for a more comprehensive output.

Researcher Focus:

Exploring the network from the perspective of a single researcher opens up many possibilities. Using an author's ORCID as an example, we can retrieve a wealth

(a) 2014 (b) 2015 (c) 2016

(d) 2017 (e) 2018 (f) 2019

(g) 2020 (h) 2021 (i) 2022

Figure 5.1: Topic-based focus: "Deep Learning" example (2014-2022).

of information related to that person, including their co-authorship, the range of publications, and their influence in the research community.

As shown in Figure 5.2, clicking on a node reveals more information about the connected nodes, such as co-authors and their respective publications. A unique feature in our network visualization is the "expand" option in the tooltip. This allows users

to take a deeper look into the co-author network, bringing co-authors of co-authors into view. Such a feature can help identify potential collaborators or new research directions, thus providing valuable insights for both the individual researcher and their broader research community.

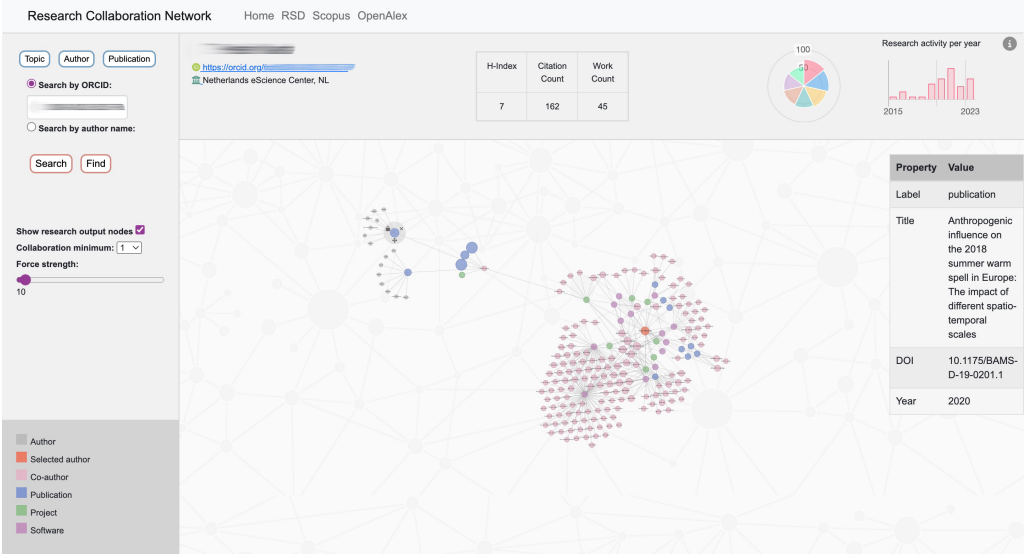


Figure 5.2: Researcher Focus: Single Author Example

Institutional Focus:

An institution's research collaboration network can be an illuminating source of information. For instance, Figure 5.3 provides a broad understanding of the research environment at the Netherlands eScience Center 2023 by exploring the institution's co-authorship network.

The network enables users to retrieve basic information about the institution, in addition to learning about the various research topics of interest to researchers at the institution. This is achieved by incorporating topic modeling into the network, which generates an overview of key research areas.

- **Research Question 2: What are the possible advantages, limitations, and biases of using network analysis to construct and evaluate collaborative research networks?**

  - 2.1. What are the possible biases and limitations of network analysis methods used to create collaborative networks?

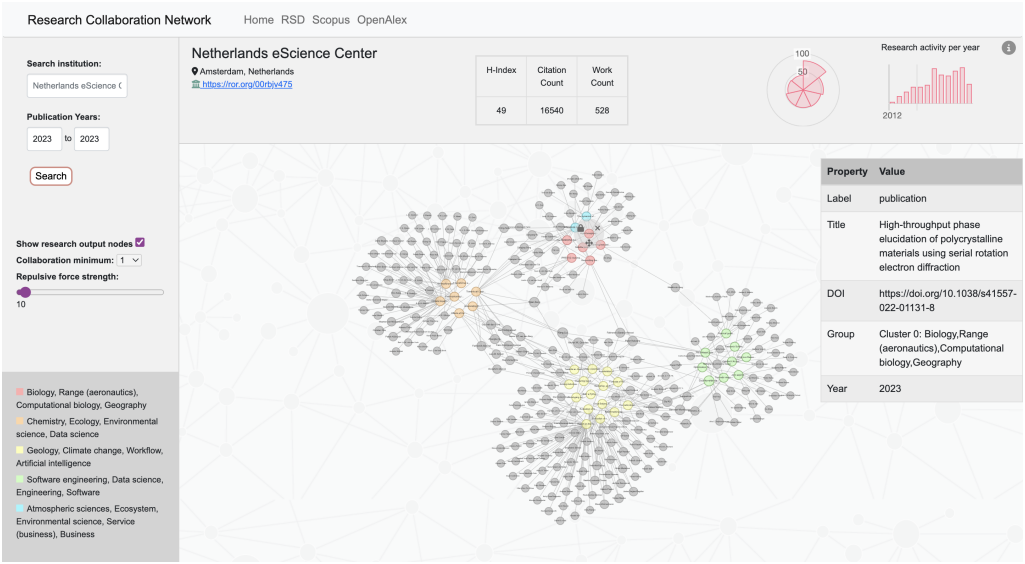


Figure 5.3: Institutional Focus: Single Institution Example

Our study has indeed exposed some limitations and potential biases, particularly related to the quality and completeness of the data, as well as the potential for algorithmic bias. These have been discussed extensively in the 'Threats to Validity' Chapter 7.

- 2.2. How can research collaboration networks improve research outcomes?

Our network of research collaborations presents many successful collaborations. By examining these collaborations, we can identify models that may be useful for future collaborations. Success can be measured in several ways: high numbers of co-authored publications, significant interdisciplinary collaborations, impactful research, or collaborations that persist over time.

By examining these successes, we can understand the composition and dynamics of effective teams. We can also identify research questions related to productive collaborations that can provide a basis for future collaborative efforts.

- 2.3. What strategies can be implemented to assess and ensure the long-term sustainability and impact of these networks?

Continuous network analysis, including regular updates of network data to take into account new collaborations, publications, and changes in researcher affiliations. In addition, network metrics need to be reassessed, to take into account new data, and

to ensure that they accurately reflect the current state of collaboration within the network. Patterns of collaboration within networks can also change over time. Periodic reassessment of these patterns is critical to keeping up with the changing landscape of research collaborations. For example, emerging collaborations may reshape the structure of the network, or the collaborative strategy of a particular research institution may change, leading to the emergence of new clusters within the network.

For long-term viability, continuous data organization is important. This includes regularly updating the network with new research results and ensuring the quality and consistency of the data fed into the network. It is also necessary to keep the network analysis methods up-to-date to increase the depth and breadth of the network.

6

Related Work

Newman (2000, 2001, 2004) (19, 48, 49) utilized static databases to construct multiple collaboration networks of researchers and examined the variations in collaboration resulting from differences in research subjects, his works are seminal in the field of scientific collaboration networks. Co-authorship in research papers was employed as the basis for determining relationships between scientists, as it represents the most intuitive form of collaboration. Newman also introduced the concept of breadth-first search to identify the shortest path between two researchers in a network model and discussed the clustering coefficient as a measure of the likelihood that two authors collaborate on a paper.

In 2004, Newman extended his research by constructing an affiliation network, which depicted connections between researchers based on their co-authorship in one or more papers. To optimize computational efficiency, the names of authors were stored in an ordered binary tree. Subsequently, names were extracted from the metadata of databases, and edges were established between pairs of co-authors on each paper. (1)

While these studies laid the foundation for constructing co-authorship networks, they lack comprehensive experimental details for comparative analysis. Our work builds upon Newman's fundamental concepts by incorporating real-time data from diverse sources such as Scopus, CrossRef, and OpenAlex. This comprehensive dataset enables us to construct not only co-authorship networks but also explore other types of collaboration networks, such as citation networks and interdisciplinary collaboration networks. Additionally, we employ advanced visualization techniques using D3.js, offering interactive and customizable visualizations that enhance the exploration and understanding of collaboration patterns and network structures. By leveraging real-time data and advanced visualization, our research provides a more comprehensive analysis of collaboration dynamics, extending

beyond the limitations of static databases. We aim to bridge the gap by providing experimental details, enabling robust comparisons and deeper insights into the complexities of research collaboration networks.

In addition to Newman's contributions, other relevant studies have focused on specific aspects of collaboration visualization. Bian *et al.* (2014) proposed CollaborationViz (4), an interactive visualization tool for biomedical research collaboration networks. These works emphasize specific dimensions of collaboration networks, such as temporal dynamics and domain-specific collaborations (Figure 6.1). In contrast, our research aims to provide a more comprehensive analysis by incorporating multiple collaboration network types, integrating real-time data, and employing advanced visualization techniques. This comprehensive approach allows us to gain insights into collaboration patterns, identify key actors and communities, and understand the evolving dynamics of scientific collaboration across diverse domains.

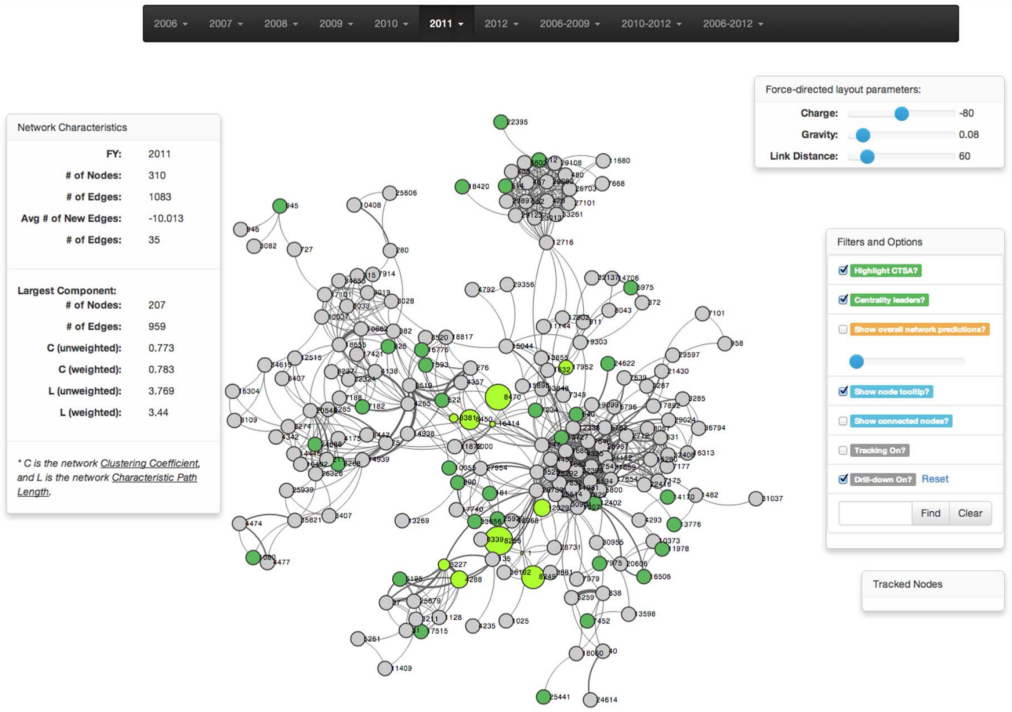


Figure 6.1: The main interface of CollaborationViz, an interactive visual analytical tool for the exploration of biomedical research collaboration networks. (4)

VOSviewer is a software tool developed by the Centre for Scientific and Technological Research at Leiden University and first launched in 2009. (50) The tool's expertise lies in the construction and visualization of scientific knowledge networks, which can be con-

structed based on citations, co-citations, or co-authorships, and is usually used for data in large bibliographic databases (e.g. Scopus, Web Science or OpenAlex). VOSviewer does not provide built-in data, but rather tools to visualize and analyze data entered into the system by the user, so it cannot merge datasets from different sources, but the user can manually integrate and clean up multiple datasets before import and analysis.

Sun *et al.* constructed a heterogeneous bibliographic network in 2011 (51). as opposed to the traditional co-author relationship. There are multiple pieces of information in the real metadata in a heterogeneous network, and the edge between each pair of authors may represent different relations. Sun *et al.* proposed PathPredict, a co-author prediction approach for a heterogeneous network. For example, two authors may be linked because they share co-authors, they may have published papers at the same conference, and so on. In this study, topological features are extracted for computing association weights, and a variety of meta paths among researchers, such as citing and cited relationship, indirect co-authorship, indirect citation relationship, and using the same citation, are examined.

In the study "Co-authorship network analysis in health research: method and potential use", Fonseca *et al.* (2016) (52) meticulously demonstrate the methodology for employing social network analysis (SNA) to examine collaboration trends and pinpoint leading scientists and organizations in the realm of health research. They present a structured approach, starting from the retrieval of scientific publications, the standardization of author and organization entries, the visualization of networks, and the calculation of metrics. The researchers apply their methodology in a case study, examining the global research network around Chikungunya virus vaccine development. Their work is predominantly methodological, providing a detailed roadmap for executing co-authorship analysis using SNA, and making use of specific software tools such as Gephi, Ucinet, and Pajek for network visualization and statistical analysis.

In the research conducted by Ghazal Kalhor *et al.*, the author delved into an analysis of the co-authorship network within Google Scholar, focusing on a unique aspect known as the Manually Added Co-authorship Network (MACN) (53). MACN, in contrast to conventional co-authorship networks that primarily depend on author lists of papers, encompasses authors who manually select their collaborators. This design renders MACN less dense but more accurate as it underscores genuine collaborations. Among the noteworthy findings of Kalhor's work is the linear correlation between an author's h-index and citation count and the discovery that the field of interest plays a crucial role in forming links between users. The work also introduces a novel citation metric, which assists in understanding the position of authors within their research area and ranking universities in each scientific field.

While the emphasis on manual collaboration and the introduction of a new citation metric are unique aspects of Kalhor's research, our work differs by focusing on the utilization of co-authorship networks for real-time collaboration detection and facilitating potential new collaborations.

Table 6.1 presents a summary of the findings of the relevant studies.

Table 6.1: Summary of Relevant Literature

Title	Author	Year	Description	Difference
The structure of scientific collaboration networks	M. E. J. Newman	2000	Investigated structure of scientific collaboration networks using mathematical models.	In addition to analyzing the structure, our study leverages both static (Scopus and RSD) and real-time (OpenAlex) data sources, providing a dynamic perspective on the scientific collaboration network within the Netherlands.
Scientific collaboration networks. I. Network construction and fundamental results	M. E. J. Newman	2001	Explored network construction principles and provided fundamental results on scientific collaboration.	Our study uses Newman's principles, but expands by utilizing a combination of graph databases and APIs, offering a detailed view of a wider range of research areas.
Co-authorship networks and patterns of scientific collaboration	M. E. J. Newman	2004	Studied patterns of scientific collaboration using a large co-authorship dataset	Our study offers a more user-oriented approach, providing interactive search capabilities for topics, authors, publications, and institutions. Additionally, it includes clustering techniques to enable users to better navigate and interpret large networks.
Software survey: VOSviewer, a computer program for bibliometric mapping	N. J. van Eck *et al.*	2009	A software tool designed for constructing and visualizing bibliometric networks based on citations, co-citations, or co-authorship relations.	VOSviewer itself does not directly combine datasets from different sources, while we offer solutions for merging different datasets.

Continued on next page

Table 6.1 – continued from previous page

Title	Author	Year	Description	Difference
Co-author Relationship Prediction in Heterogeneous Bibliographic Networks	Y. Sun *et al.*	2011	Proposed a framework to predict co-author relationships in heterogeneous bibliographic networks.	Our work, while also involving co-author networks, is more focused on interactive visual exploration and does not include predictive aspects.
CollaborationViz: Interactive Visual Exploration of Biomedical Research Collaboration Networks	J. Bian *et al.*	2014	Developed a tool specifically for the interactive visualization of biomedical research collaboration networks.	While our work also provides interactive visualizations, it does not focus on a specific research field and additionally offers dynamic data acquisition and topic, author, publication, and institution search capabilities.
Collaboration Map: Visualizing Temporal Dynamics of Small Group Collaboration	S. Lim *et al.*	2015	Created a visualization tool to analyze the temporal dynamics of small group collaborations.	Our work also visualizes collaborations, but it encompasses a broader scale and incorporates a wider range of research areas, also leveraging dynamic and static data sources.
Topic Modeling of Document Metadata for Visualizing Collaborations over Time	F. Chen *et al.*	2016	Implemented topic modeling to visualize collaborations over time based on document metadata.	Our study also provides a visualization of collaborations, but goes beyond by including search capabilities, real-time data acquisition, and clustering techniques to handle large networks.
Co-authorship network analysis in health research: Method and potential use	B. P. Fonseca *et al.*	2016	Reviewed the method and potential applications of co-authorship network analysis in health research, and provided an example.	Our work does not limit its scope to health research, and provides a more interactive, user-oriented tool with extensive search functionalities.
A new insight into the analysis of co-authorship in Google Scholar	G. Kalhor *et al.*	2022	Investigated the MACN of Google Scholar, with a focus on authorship collaboration, and introduced the structural properties of MACN and a new citation metric (MCC-index).	This work uses manual co-authorship data from Google Scholar, and focuses on individual author characteristics, while our work focuses on real-time data usage, interactivity, and user-based search functionality.

7

Threats To Validity

In this section, we critically evaluate potential threats to better understand the extent to which we can trust the application and findings derived from the study, according to the classification framework proposed by Wohlin *et al.* (54).

7.1 Internal Validity

Internal validity refers to the integrity of the causal relationships that have been inferred within the scope of the study, it is the validity of the conclusions drawn from the data in the research setting. (54). Within the context of our research on co-authorship networks using Scopus, OpenAlex, and RSD data, several factors could potentially threaten the internal validity. These include but are not limited to, the quality of the data we used, the potential biases of our algorithms and analytical methods, and the limitations of the database and visualization tools.

7.1.1 Data Quality

The quality and completeness of data used in this study pose significant threats to its internal validity. Several limitations in our key data sources, Scopus, OpenAlex, and RSD could potentially compromise the inferences drawn from our analysis.

The lack of a consistent unique identifier across all databases is a major challenge. ORCID serves as an essential unique identifier for authors, and DOI for publications. Unfortunately, Scopus uses its own Scopus ID and does not provide ORCID information in the standard service. On the other hand, OpenAlex does provide ORCID data, but a significant portion of this data is missing, rendering it less than ideal for merging with

other data sources. Similarly, Research Software Directory, despite providing comprehensive study data for Dutch institutions, including unpublished projects and software, suffers from a high proportion of null values for researchers' ORCIDs. This impedes the distinction between different researchers and hinders merging RSD data with other data sources and requires translation each time new data is stored, adding time and complexity to the data management process. This identifier translation involves frequent requests to Scopus or ORCID APIs, slowing the overall system performance.

Accessibility to data is another limitation. Scopus data is only accessible via an educational VPN, which restricts access for researchers without this facility.

Moreover, the pybliometrics Python package provided by Scopus has a strict limit of 5,000 item results per week. For large-scale projects like ours, this limit is highly restrictive (38). The Scopus data we currently use is a large, statically downloaded database of metadata, meaning it will not be updated, which is a significant limitation for any web application with a practical aim.

7.1.2 Algorithmic Bias

In the process of topic modeling, we encountered several important issues that could affect our results. Initially, we attempted to classify topics using textual data from various sources, including titles, abstracts, descriptions, and keywords. Despite implementing rigorous data cleaning, stemming, and lemmatization procedures, the resulting cluster names were unclear and difficult to interpret. To solve this problem, we chose to use the Concepts List property of OpenAlex. However, since these concept words are relatively fixed, this solution leads to a relatively rigid classification of topics.

Another challenge is the computational inefficiency of our algorithm. Since we use OpenAlex data to classify each publication by topic, we need to send a large number of requests to retrieve keywords for different papers and projects. This significantly slows down the program, which can be a major obstacle for large-scale applications or applications that require real-time analysis.

7.1.3 Database and Visualization Limitations

While Neo4j is a powerful tool for storing and querying graph data, it is not without its limitations. A major problem is its high memory requirements for large graphs. When a database consists of a large number of nodes and links, the performance of the database creation and query functions may degrade, especially if unique identifiers are not available.

If unique identifiers are provided, they can be used as constraints on nodes or links, acting as indexes and thus reducing computational costs.

In addition, Neo4j relies on a single model (attribute graph model) to represent the data. This may not perfectly encapsulate the complexity inherent in the ensemble network (41). In addition, data processing to extract nodes and links from the database based on query criteria is a challenging task that complicates the retrieval of complex network graphs.

Finally, the cloud storage capabilities and data sharing options offered by Neo4j are more costly than traditional databases. These challenges underscore the need to carefully interpret our results, and explore alternative or complementary databases and methods for future research.

7.2 External Validity

External Validity relates to the extent to which the findings of this research can be generalized or applied to other contexts, settings, or groups (55). Several factors within our study may influence the external validity of our results.

7.2.1 Data Scope

The range of data used in this study may limit its external validity. The data sources we relied on, Scopus, OpenAlex, and RSD, while comprehensive, each have their areas of focus and limitations.

For real-time network generation, we utilize OpenAlex's real-time API, considering its minimal limitations and speed, but this system operates separately from Scopus and RSD's datasets. Therefore, our results may not be fully representative or generalizable to other academic fields, different databases, or co-authored networks from different periods. The inclusion of other metadata sources has the potential to address some of the issues associated with academic fields or different databases. However, the challenge of reconciling the different unique identifiers used by the various databases remains.

For example, co-authorship patterns and network characteristics found in other disciplines may differ from our study due to differences in research domains, collaborative practices, and publication norms. In addition, our results may be influenced by the specificity of the databases we use. The results may differ if data from other databases are used for the analysis. In particular, the completeness and accuracy of records vary greatly across data sources. For example, a query of a particular author in Google Scholar may

yield 47 research outputs, while the same author may be associated with only 12 publications in CrossRef and Scopus. The reason for this difference is that Google Scholar may include unpublished works or incorrect information, while CrossRef and Scopus databases may lack some records.

7.2.2 Tools and Techniques

The specific tools and techniques used in this study may also influence our findings. We used Neo4j for database management, D3 for visualization, and topic modeling for data analysis. These choices were based on our specific research needs. However, if this study were replicated with a different author identification method or clustering algorithm, it is possible that the results would be different. For example, different visualization tools, such as Cytoscape, which provides additional community selection features, may show different visual interpretations. In addition, the topic modeling algorithms we used may affect the results of topic clustering.

7.3 Construct Validity

Construct validity refers to the extent to which the tools or measures used accurately capture the constructs they are intended to measure. In our study, there are two key constructs, namely the representation of the study topic through thematic modeling and the description of co-authors in our chosen database.

7.3.1 Topic Clustering

We adopted n-gram and topic modeling approaches to cluster research topics, but these techniques, despite being robust, are not infallible. Topic modeling assumes that words used in the same context carry similar meanings, but this assumption may not always hold. For example, words such as 'model', 'project', 'tool', etc., may appear frequently across various topic groups. Although these words have valid meanings, they do not contribute to a clear distinction among topics. In addition, the same terms can be used differently across disciplines. A case in point is "residual risk", a term usually used in the Economic and Risk context, which showed up in a Medical paper where it referred to residual cardiovascular risk (56). Similarly, the n-gram approach, while efficient, may fail to capture broader themes that span multiple phrases or sentences.

These limitations might impact our ability to accurately represent the actual topics of the articles. Moreover, the arbitrary nature of determining the number of clusters in topic

modeling can result in overfitting or underfitting, meaning that the true nature of the co-authorship network may not be accurately represented.

7.3.2 Representation of co-authorship

The representation of co-authorship relationships in our database is another important construct. While our database captures certain types of collaborations, it may not fully encompass the breadth of co-authorship relationships that exist in academia. For example, it may not accurately represent multidisciplinary collaborations, collaborations between different institutions or countries, or non-traditional collaborations, such as those between academics and industry professionals.

In addition, it is worth noting that authorship and contributions are complex structures that are not always fully reflected by co-authorship links. For example, an author's position in the author list may have different meanings in different disciplines, and some substantive contributions may not result in co-authorship.

7.4 Conclusion Validity

Conclusion Validity is concerned with the reliability of the inferences made about the relationships among variables based on the collected data. Given the complex visualizations generated from large datasets in our study, there is potential for bias or error in the interpretation of the data and results. The interpretation of network structure, cluster associations, and patterns may be subjective, and different node relationships and distributions may vary between observers. In addition, complex visualizations may inadvertently obscure important details or relationships in the data, leading to potential misinterpretations. Therefore, we should be cautious in interpreting visual results and drawing conclusions based on them.

8

Conclusion

8.1 Summary of Objectives and Findings

We present an integrated framework for constructing and analyzing these networks, demonstrating the utility of integrating data from a variety of sources and the effectiveness of advanced network analysis techniques. The primary purpose of this study is to investigate the creation, analysis, and repercussions of research collaboration networks, with a focus on co-authorship networks, which pave the way for a deeper understanding of the dynamics of collaborative research. Our investigation shows that meticulous data collection, cleaning, and integration are the basis for building robust collaborative networks. Our investigation reveals trends, patterns of collaboration, and intricate connections among researchers and institutions, providing valuable experience for the field of collaborative research.

Studying collaborative networks has important theoretical and practical implications. From a theoretical perspective, we confirm the utility and validity of using collaborative networks as a powerful tool for analyzing and understanding research collaboration. In terms of practical application, the results are useful for strategic planning of future research collaborations. Our work can facilitate the identification of potential collaborators and thus facilitate interdisciplinary research collaborations.

8.2 Future Research

Drawing from the lessons learned and the experiences gained in this study, we suggest the following potential directions for future research:

- **Broadening Data Source Usage:** While this study primarily hinged on data from Scopus, OpenAlex, and RSD, future studies could widen the scope by incorporating

more data sources to capture a fuller picture of research collaborations.

- **Counteracting Biases:** There are potential biases in our study, such as those introduced by topic modeling and co-authorship representation. Future research should strive to address and further reduce these biases.
- **Sustaining Networks in the Long Run:** The importance of ongoing evaluation and updating of collaborative networks is an important takeaway from our study. Future research should develop strategies in promoting the longevity of these networks.
- **Applying the Findings:** Future work could focus on planning and managing research collaborations to be utilized in real-world scenarios.

In conclusion, this study has greatly enriched the understanding of research collaboration networks and hopefully, it will inspire exploration and breakthroughs in this field.

References

[1] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5200–5205, 2004. v, 12, 39

[2] Mohamad Alnajem, Mohamed M. Mostafa, and Ahmed R ElMelegy. Mapping the first decade of circular economy research: a bibliometric network analysis. *Journal of Industrial and Production Engineering*, 38(1):29–50, 2021. v, 12

[3] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents, 2012. v, 14

[4] Jiang Bian, Mengjun Xie, Teresa Hudson, Hari Eswaran, Mathias Brochhausen, Josh Hanna, and William Hogan. Collaborationviz: Interactive visual exploration of biomedical research collaboration networks. *PloS one*, 9:e111928, 11 2014. v, 40

[5] Sooho Lee and Barry Bozeman. The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35(5):673–702, 2005. 1

[6] Caroline S. Wagner and Loet Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10):1608–1618, 2005. 1

[7] Wolfgang Glänzel and András Schubert. *Analysing Scientific Networks Through Co-Authorship*, pages 257–276. Springer Netherlands, Dordrecht, 2005. 1, 11

[8] Sooho Lee and Barry Bozeman. The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35(5):673–702, 2005. 1

[9] Carter Butts. Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, 11:13 – 41, 03 2008. 2

[10] Evelien Otte and Ronald Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6):441–453, 2002. 2, 7

[11] Katy Borner, Chaomei Chen, and Kevin Boyack. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37:179–255, 01 2005. 3

[12] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978. 7

[13] Linton C Freeman, Cynthia M Webster, and Deirdre M Kirke. Exploring social structure using dynamic three-dimensional color images. *Social Networks*, 20(2):109–118, 1998. 7

[14] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, 1994. 7

[15] Stan Wasserman and Philippa Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrika*, 61:401–425, 09 1996. 7

[16] Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009. 7

[17] Peter J Carrington, John Scott, and Stanley Wasserman. *Models and methods in social network analysis*, volume 28. Cambridge university press, 2005. 8

[18] Basic SNA concepts — ibm.com. `https://www.ibm.com/docs/en/csfdcd/7.1?topic=concepts-basic-sna`. [Accessed 22-Jun-2023]. 9

[19] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, Jun 2001. 9, 39

[20] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010. 9

[21] Alireza Abbasi and Jorn Altmann. On the correlation between research performance and social network analysis measures applied to research collaboration networks. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10, 2011. 9

[22] Loet Leydesdorff. "betweenness centrality" as an indicator of the "interdisciplinarity" of scientific journals. *Journal of the American Society for Information Science and Technology*, 58, 07 2007. 9

[23] M. Salter-Townshend, A. White, I. Gollini, and T. B. Murphy. Review of statistical network analysis: models, algorithms, and software. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(4):243–264, 2012. 10

[24] J. MacQueen. Some methods for classification and analysis of multivariate observations. 1967. 10

[25] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010. 11

[26] Loet Leydesdorff and Caroline S. Wagner. International collaboration in science and the formation of a core group. *Journal of Informetrics*, 2(4):317–325, 2008. 11, 13

[27] Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965. 11

[28] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265 – 269, 07 1973. 12

[29] Lee Fleming and Olav Sorenson. Technology as a complex adaptive system: evidence from patent data. *Research Policy*, 30(7):1019–1039, 2001. 13

[30] Ian Marques Porto Linares, Alex Fabianne De Paulo, and Geciane Silveira Porto. Patent-based network analysis to understand technological innovation pathways and trends. *Technology in Society*, 59:101134, 2019. 13

[31] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001. 13

[32] James Moody. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2):213–238, 2004. 13

[33] Diane Sonnenwald. Scientific collaboration. *Annual Review Of Information Science And Technology*, 41:643–681, 12 2007. 13

REFERENCES

[34] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003. 13

[35] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 990–998, New York, NY, USA, 2008. Association for Computing Machinery. 14

[36] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 306–315, New York, NY, USA, 2004. Association for Computing Machinery. 14

[37] Qi Wang and Ludo Waltman. Large-scale analysis of the accuracy of the journal classification systems of web of science and scopus. *Journal of Informetrics*, 10(2):347–364, 2016. 17

[38] Elsevier. How much data can i retrieve with my apikey? 17, 46

[39] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. 18

[40] Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Comput. Surv.*, 40(1), feb 2008. 21

[41] Jim Webber. A programmatic introduction to neo4j. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, SPLASH '12, page 217–218, New York, NY, USA, 2012. Association for Computing Machinery. 21, 47

[42] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media, Inc., 2nd edition, 2015. 21

[43] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009. 22

[44] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003. 22

[45] Mike Bostock. D3.js - data-driven documents, 2012. 22

[46] Neo4j browser user interface guide. `https://neo4j.com/developer/neo4j-browser/`, Last accessed on 2023-06-30. 24

[47] Mozilla Developer Network. XMLHttpRequest. `https://developer.mozilla.org/en-US/docs/Web/API/XMLHttpRequest`. Accessed on: 02 06 2023. 25

[48] M. E. J. Newman. Models of the small world: A review, 2000. 39

[49] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64:016131, Jun 2001. 39

[50] Nees Jan van Eck and Ludo Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84:523–538, 08 2010. 40

[51] Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal, and Jiawei Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 121–128, 2011. 41

[52] Bruna de Fonseca, Ricardo Barros Sampaio, Marcus Vinicius Fonseca, and Fabio Zicker. Co-authorship network analysis in health research: Method and potential use. *Health Research Policy and Systems*, 14(1), 2016. 41

[53] Ghazal Kalhor, Amin Asadi Sarijalou, Niloofar Sharifi Sadr, and Behnam Bahrak. A new insight to the analysis of co-authorship in google scholar. *Applied Network Science*, 7(1), 2022. 41

[54] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in Software Engineering - An Introduction.* Kluwer Academic Publishers, 2012. 45

[55] Chittaranjan Andrade. Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian Journal of Psychological Medicine*, 40:498 – 499, 2018. 47

[56] Henry N Ginsberg, Chris J Packard, M John Chapman, Jan Borén, Carlos A Aguilar-Salinas, Maurizio Averna, Brian A Ference, Daniel Gaudet, Robert A Hegele, Sander Kersten, Gary F Lewis, Alice H Lichtenstein, Philippe Moulin, Børge G Nordestgaard, Alan T Remaley, Bart Staels, Erik S G Stroes, Marja-Riitta Taskinen, Lale S Tokgözoğlu, Anne Tybjaerg-Hansen, Jane K Stock, and Alberico L Catapano. Triglyceride-rich lipoproteins and their remnants: metabolic insights, role in atherosclerotic cardiovascular disease, and emerging therapeutic strategies—a consensus statement from the European Atherosclerosis Society. *European Heart Journal*, 42(47):4791–4806, 09 2021. 48