Social Network Analysis Applied to Research Collaboration: A Literature Review

Zhining Bai[12]
[1]Faculty of Science, Vrije Universiteit Amsterdam, the Netherlands
[2]Faculty of Science, Universiteit van Amsterdam, the Netherlands
zhining.bai@student.uva.nl
2728339(VU), 13714538(UvA)

Abstract

This paper presents a review of papers on the study of social network analysis (SNA) in academic research collaboration scenarios, dismantling and summarizing the methods and tools used in previous related research at different stages of the network analysis process. In this literature study, the goal-question-metric (GQM) approach to literature reading was used to orient the literature study, providing an overview and summary of techniques in collaborative networks such as metadata acquisition, data pre-processing, topic modelling, social network analysis, and visualization techniques.

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 [24]. Cross-disciplinary directions such as Artificial Intelligence (neural networks), Bioinformatics, and Nanoscience, as well as the advancement of traditional disciplines, promote multidisciplinary cross-collaboration. Therefore, finding potential research collaborators and expanding collaboration networks is critical in the development of disciplinary research. Netherlands eScience Center has an extensive track record of collaborative projects with researchers from all over the Netherlands (and beyond), but it is difficult to know whether we have approached all potential collaborators. Creating a collaboration network of researchers all over the Netherlands allows analysis of historical collaboration information to predict the propensity of researchers to collaborate, and help us identify collaboration blind spots, such as researchers and groups who have never worked with us, and whether there is a mismatch between researchers we are actively collaborating with. To this end, I will create a researcher exchange network within the Netherlands to explore collaboration information from published research in recent years through clustering techniques and network analysis techniques, and combine it with our past research collaborations to develop interactive, force-directed network maps.

In general, social network analysis builds social networks based on records of information exchange among members of an organization or company. Research collaboration networks are a subset of SNA in which co-author information from academic journal papers is used to build scientists' research collaboration social networks, but there are studies that suggest that co-authorship is only a partial indicator of research collaboration [22], and citations can also represent communication activity between researchers. In addition to publication collaborations, there are a variety of important collaborative researches that had not been submitted for publication.

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 [10]. 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 [32]. The analysis of research collaboration networks is therefore more complex.

The social network analysis (SNA) of research collaboration will be the main focus of this article, which will review and analyze the relavant literature. My study on researchers' social network analysis will extensively discuss co-authorship and citation networks in publications and other networks of scientific interactions.As for collaboration networks in scientific research, the connections between researchers are mainly reflected in the literature of research results, co-authorship and citation networks will be the focus of discussion.

The structure of this essay consists of four sections. The first section covers the paper's structure, overall subject matter, and an introduction to the SNA theory. A thorough investigation into social networks and research networks is also described. The methodology of the study is presented in Section 2, which is divided into three subsections: the research objectives, the definition of the five research questions and how they served to guide the literature review, the acquisition and integration of the literature, and the experimental design, which was chosen using the goal-question-metric (GQM) approach. The researcher's SNA is examined in Section 3 in accordance with a subsection order that follows the social network research workflow. Section 4 summarises the research methodology of the paper discussing co-authorship and citation networks, and the solution to the research question, and the final section concludes with a summary of the current state of research on researchers' social networks, a summary of the paper's work and an outlook on the research to follow.

2 Study Design

This section's design will guide this literature study, and the research goals, research questions, and literature classification and integration metrics are presented below using the Goal-Question-Metric approach [23].

2.1 Research Goal

To fully characterize the research collaboration network design, the research goal is presented in Table 1.

Table 1: Research Goal

Purpose	Create national collaborative research networks in the Netherlands to promote collaboration among researchers.
Issue	A multiple force-oriented graph presenting a collaboration network of researchers is created by obtaining personal information and collaborative communication data (co-authorship, citation relationships, other communication relationships) of researchers, with nodes being researchers with similar research classified through clustering techniques such as topic modelling, and the force of the edge between two nodes calculated according to the collaborative relationship.
Object	Researchers with published projects and papers (in English) who have been active in recent years all over the Netherlands.
Viewpoint	In the literature of similar research areas, clusters of cooperative networks can be presented between researchers in close geographical proximity.

2.2 Research Questions

The following research questions and descriptions were used to guide the direction of the research in the literature review.

RQ 1: What are the primary methods for obtaining accurate and up-to-date metadata for a collection of papers in the relevant literature? What are the most important metadata parameters to pay attention to?

- Obtaining all the metadata of the literature in real time is a huge challenge, and the large amount of redundant data can cause computational crashes. Also, many papers are located in different publishers and therefore cannot be accessed simultaneously. The literature search APIs of different institutions often provide parametric filters for narrowing down result sets. An efficient method of data extraction is very important in this study.

RQ 2: How to filter the data to study only researchers in the Netherlands?

- In this research, we put the perspective of researchers located in the Netherlands, but it is difficult to collect metadata directly because there is no such geographically based database, especially one that covers all researchers in a given field. An efficient method of pre-processing metadata would therefore ensure the accuracy of the scope of our study.

RQ 3: What is the role of topic modelling in the overall workflow and what advanced technologies are currently available that could be useful for our research?

- Creating a collaboration network of authors demonstrates how researchers relate to one another, and it is also important to know what they are working on. Using topic modeling allows for the clustering of related studies, which helps avoid the errors and noise associated with constructing datasets solely based on keyword searches.

RQ 4: How to define the strength of relationships between researchers in a force-directed graph?

- In a force-directed graph, the relationships between nodes are influenced by attractive forces and repulsive forces, and how to calculate the strength of the forces in the relationship between the nodes needs to be confirmed.

RQ 5: What are some common tools for visualising social networks?

- The purpose of this question is to identify visualisation tools available from previous research that demonstrate collaboration networks, and to provide visualisation ideas for the following research.

2.3 Literature Extraction and Synthesis

The papers in this literature review were sourced from Google Scholar, IEEE Explore, ACM Digital Library, Springer Link, Elservier Scopus, and other search engines, using keywords such as "social network analysis" and "collaboration networks" to find relevant studies and expand the literature collection by exploring their citations using the snowball approach [21]. Papers will be classified into three categories based on collaborative research networks: co-author networks, citation networks, and other developer networks. Using the SNA research steps, the methods used in the relevant studies will be summarized sequentially.

3 Literature Review

3.1 Data Extraction and Data Pre-processing

In this section, I will summarize the data extraction and pre-processing methods found in the relevant literature. I will look for the most effective and popular methods of obtaining researcher metadata by systematically collating the papers in order to inspire my own research.

Since SNA focuses on network structure rather than individual behaviors, all relationships within a closed population should be gathered. The data needed for analysis and calculation must cover explicit or hidden relationships, i.e., correlated data that can connect nodes into a network [27]. Co-authorship and citation relationships in the literature metadata can provide obvious data on social behavior for collaboration and communication networks of researchers.

For the stable generation of networks, bibliographic database is the first choice for the previous research. Newman (2000, 2001, 2004a) [28, 29, 30] studied co-authorship networks in specific disciplines such as physics, biology, and mathematics, using datasets such as the open access paper databases MEDLINE (bibliographical database from 1995 to 1999), Networked Computer

Science Technical Reference Library (NCSTRL), and Stanford Public Information Retrieval System (SPIRES). The sheer size of the database makes it difficult to accurately screen the literature database for duplicate authors. Since an author may use different name abbreviations in different papers, or there may be a number of authors with the same name, and the affiliations and region may change for each researcher, Newman used two solutions for comparison, the first is to consider only the surname and the initials of the name, and the second is by surname and initials. Both methods effectively avoid identifying the same author as someone else, but they also tend to conflate different people. The author's unique ID data is therefore very important in the data processing process.

Otte and Rousseau (2002) [32] consulted three databases: CSA Sociological Abstracts Database (SA), Medline Advanced and PsycINFO, to construct and summarise the SNA co-authorship network. Kajikawa et al. (2007) [20] collected citation data for those publications from the Science Citation Index (SCI) and the Social Sciences Citation Index (SSCI) compiled by the Institute for Scientific Information (ISI). Abbasi and Altmann (2011)[2] proposed AcaSoNet, which is a Web-based application for extracting publication info. In addition, DBLP, a search service providing scientific literature in the field of computing, which offers open source journals and proceedings for download, is also one of the popular sources for obtaining papers in the field of computer science. For instance, in 2008, Santamaría and Therón created the overlapping visualisation tool using the dblp datase; Sun et al. (2011) [40] explored the network of co-authors of the DBLP bibliography; Lim and Chiu [26] used a small dataset in DBLP which contains publications by 29 researchers at their lab for 25 years to build a collaboration map. Chen et al. (2016) [11] used the InfoVis dataset, as well as 21 years of publication metadata from their lab, to model the themes of the collaborative documents.

For literature metadata acquisition and pre-processing, Heldens et al. (2022) [18] proposed a Python package, litstudy, which supports access to papers from four resources: Scopus, Smantic Scholar, CrossRef and dblp. litstudy provides convenient query functions to find metadata from online databases. In addition, some publisher websites offer to download metadata and export to csv files, but due to the non-real time nature and small volume of data, this cannot be used as the primary method of accessing metadata for papers. Due to the different data sources, litstudy offers features such as merging and filtering document datasets to generate DocumentSet. litstudy offers customisable options for data pre-processing such as stemming, word removal, n-gram detection, etc.

Currently, bibliographic indexing APIs are all able to provide basic metadata such as title, author, publisher, publication date, DOI, etc. Some metadata search APIs also provide keyword and abstract data, which can help us with subject analysis and clustering. To distinguish between authors, some services provide author ids to avoid being affected by duplicate names and different abbreviations. Table 3 summarises the comparison of information from some common data source APIs based on the data required for this study.

Table 2: Description of the metadata provided by different literature databases

			Metadata						
Name	Discipline	Access	Author	AuthorID	Keyword	Country	Affiliation	Abstract	Citation
Scopus	Multi	Subscription	✓	✓	✓	✓	✓	✓	✓
Springer	Multi	Free	✓		✓	✓		✓	
dblp	CS	Free	✓	✓					
IEEE	CS, EE	Subscription	✓		✓		✓	✓	
CrossRef	Multi	Free	✓				✓	✓	✓
Semantic Scholar	Multi	Free	✓	✓			✓	✓	✓

In addition to co-authorship and citation networks for papers, developer collaboration networks for software development and patents can also provide data for the construction of a researcher's SNA.

Gloor et al. (2003) [15] visualised the Collaborative Innovation Network (COIN) of the W3C (WWW Consortium) working groups, where the author analysed electronic interaction logs such as emails in order to find COIN within the organisation by using Google to find relevant documents based on link patterns. Prato et al. (2012) [13] constructed a global technology collaboration network based on international co-invention data for patents, from the European Patent Office (EPO) Global Patent Statistics Database 2010. In 2013, Jermakovics et al. [19] explored the collaboration network of open-source projects from phpMyAdmin, Eclipse Data Tools Platform and Gnu Compiler Collection. Because the biomedical research collaboration networks is different from publication co-authorship, Bian et al. (2014) [6] used the metadata of research grant obtained from the Office for Research and

Sponsored Programs (ORSP) to track the researchers' roles on each grant. Access to this type of data is diverse and trivial, and it is difficult to have a uniform standard, so only a general overview is given in this paper.

3.2 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 modelling 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 modelling the author's topic of the researcher.

Probabilistic Author-Topic Models (Steyvers et al., 2004) [39] is an unsupervised learning model. It can be used to obtain information in large collections of texts and then explore the probability of combining authors with the topics of the texts. The document topics associated with each author are typically multiple, and the topics of documents with multiple authors can be seen as a combination of the related topics of their authors. The process of author topic modelling is to select one of the multiple topics of an author and then match each word in the text with the related words of that topic. Steyvers et al. used Gibbs sampling, a Markov chain Monte Carlo (MCMC) algorithm that samples from the posterior distribution of parameters to estimate parameters for authors θ and topics ϕ. The authors then applied the model to the CiteSeer digital library dataset. the topic and author assignment are sampled from the following formula, where $z_i = j$ and $x_i = k$ denote the author and topic corresponding to the ith word in the text, $w_i = m$ represents the m_{th} word in the lexicon, and z_{-i} and x_{-i} represent the historical topic and author assignments in addition to the i_{th} word.

$$P\left(z_i = j, x_i = k \mid w_i = m, \mathbf{z}_{-i}, \mathbf{x}_{-i}\right) \propto \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \frac{C_{kj}^{AT} + \alpha}{\sum_{j'} C_{kj'}^{AT} + T\alpha} \tag{1}$$

Latent Dirichlet Allocation (LDA) [7] is a way of estimating the likelihood of data using a prior distribution and ultimately obtaining a posterior distribution based on a Bayesian algorithm model. Rosen-Zvi et al. (2012) [34] proposed an LDA-based model for document collection generation, the Author-Topic Model. This can be done by modelling 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, and the author constructs a Markov chain using Gibbs sampling to estimate the parameters and random variables by sample, as shown in the formula below, a_d represents the set of authors.

$$P\left(z_i = j, x_i = k \mid w_i = m, \mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w}_{-i}, \mathbf{a}_d\right) \propto \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \frac{C_{kj}^{AT} + \alpha}{\sum_{j'} C_{kj'}^{AT} + T\alpha} \tag{2}$$

Chen et al. (2016) [11] used LDA-based topic modelling on document metadata to create visual collaboration networks that vary with different time spans. In the LDA model, each document is a mixture of topic sets rather than individual topics, thus enabling better modelling of thesis topics. k denotes the number of topics, and the set of document D represented as term vectors in V. The authors used the Gensim Python software library to compute the parameters and to downscale the learning model matrix $k \times V$, retaining only m (m=10) high probability related terms per topic. To avoid overlapping of topics, authors created a subspace V_c, with $dim(V_c) <= k * m$, to make it more efficient.

Lim et al. (2016) [25] proposed a combined author-citation network model, Citation-Network Topic Model (CNTM), using a combination of a Poisson mixed topic linking model and an author-topic model to analyse information about the author's research area, citation network and paper content. The model is a non-parametric extension of the LDA model and uses the Griffiths-Engen-McCloskey (GEM) distribution to compute probability vectors. CNTM is based on the Author-Topic Modelling approach mentioned above, using the GEM distribution to sample the root topic distribution μ, the author distribution v and the topic distribution θ. They also proposed a Metropolis-Hastings (MH)

Algorithm for citation network using Poisson Distribution, which has improved performance on both fitting and clustering tasks. For each citation x_{ij}:

$$x_{ij}, y_{ij} = k \mid \lambda, \theta' \sim Poisson(\lambda_i^+ \lambda_j^- \lambda_k^\tau \theta'_{ik} \theta'_{jk}) \tag{3}$$

3.3 Social Network Analysis

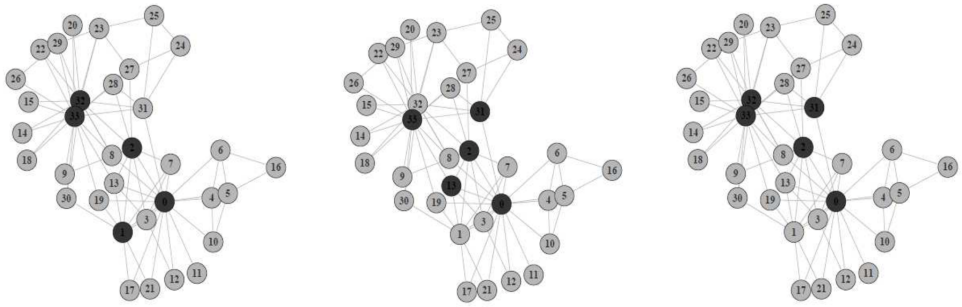
Social Network Analysis (SNA) combines techniques from social science, information science, graph theory and more. This section introduces some concepts in graph theory in SNA, metrics and some graph theory model structures for collaboration networks.

Social network graphs are divided into directed and undirected graphs, which consist of a set of nodes (social individuals) and a set of links (social relationships). In SNA we can define the graph to be complete if each node has direct connections to other nodes, and the graph density of SNA is calculated by dividing the number of links in the complete graph by the number of nodes [32]. The density D of an undirected graph G with N nodes is calculated as:

$$D = \frac{2 * (\#L(G))}{N(N-1)} \tag{4}$$

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

For node activity in a network, **Degree Centrality** assesses the activity metrics of participants, i.e. the number of direct relationships a node has with other nodes. In a common way of thinking, the number of direct connections a node has can represent its level of importance in the network. But the reality is more complex, and it is important to know whether the node is connected to people around it who are already connected to each other or whether it is exposed to new groups (Abbasi and Altmann, 2011)[2]. **Closeness Centrality** represents how close a node is to other nodes, and whether the direct and indirect connections between nodes provide a shortest path that allows quick access to other nodes in the network. **Betweenness Centrality** is another way of measuring the influence of social nodes, which takes into account the node's role as a connector among the network, where a node resides in the network when it acts as a mandatory path for two other nodes to connect, if a single point of failure would cause a blockage in the entire network (Salter-Townshend, 2012)[36].



(a) Degree Centrality (b) Closeness Centrality (c) Betweenness Centrality

Figure 1: Degree, closeness and betweenness centrality measures of a case network, where the black nodes are the ones with the highest centrality scores. [36]

3.4 Social Network Analysis for Research Collaboration

3.4.1 Co-authorship Network

Newman (2000, 2001, 2004a) [28, 29, 30] used static databases, as summarised in Section 3.1, to construct multiple collaboration networks of researchers, and explored the differences in collaboration

caused by differences in research subjects. The relations between scientists were determined by their co-authorship in the research papers, as this is the most intuitive collaboration. Newman also discusses breadth-first search to find the shortest path between two researchers in a network model, and clustering coefficient that represents the possibility that two authors coauthor in one paper. In 2004, Newman constructed affiliation network which represented the links between two researchers by co-authorship in one or more papers. The names of authors are stored in an ordered binary tree to reduce the computing cost, and then names are extracted from the metadata of databases, and edges are drawn between pairs of coauthors on each paper. These studies provide the fundamental concept of co-authorship network construction but lack experimental details for comparison.

Otte and Rousseau(2002) [32] created an undirected co-authorship network graph using 1601 articles containing 133 authors, 57 of whom led to a big associated circle. The undirected graph is chosen because the co-authorship relationship is bi-directional. The main network analysis algorithm they used is UCINET [8], which provides relation analysis and centrality calculation. In terms of the cliques, we can summarize the frequency of authors' publication by Lotka's Law [35], it can be described as the following equation:

$$f(y) = \frac{0.790}{y^{2.727}} \tag{5}$$

where $f(y)$ the number of authors with y articles. Therefore, there are a large amount of singletons and small association composed of two authors in the network.

Sun et al. constructed a heterogeneous bibliographic network in 2011 [40]. as opposed to the traditional co-author relationship. There are multiple 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.

3.4.2 Citation Network

Citation networks are another way of identifying relationships between authors, and by analysing citation networks we can understand the structure of a field of study. Unlike co-authorship networks, because the papers cited and citing are distinct, the relationships in a citation network are directional.

Kajikawa et al. (2007) [20] created a scholarly citation network on the topic of sustainability science using a topological clustering approach to classify sustainability science into 15 clusters. As there may be some citation papers that are not related to sustainability science, the authors focused on the maximum connected component to analyse the sustainability science structure. The model analysis procedure consists of three major steps: First, the citation data are assembled into a directed graph comprised of components and single nodes. The network is then converted into a non-direct graph with no weights. Finally, using a clustering algorithm, the citation network is divided into clusters (Newman, 2004b) [31]. The clustered networks were arranged more compactly by using a spring layout algorithm for papers citing each other.

In a co-citation network, paper are said to be co-cited if both cited by another paper. When authors or papers are co-cited frequently, it can be seen that they have similar research ideas and hence belong to a similar research cluster. Author co-citation analysis (ACA) can demonstrate the collaboration of researchers' work as well as the evolution of the discipline. Bu et al. (2016) [9] proposed a modified author co-citation analysis (MACA) method based on ACA, which builds an author citation network using citation metadata such as conference, publication time, place, keywords, and publication vehicle. Based on the four metadata of author, time, carrier and keyword, the relationship coefficients of research fields of two co-cited authors can be calculated.

In 2021, Alnajem et al. [4] provided a bibliometric analysis of circular economics (CE) from 2009 to 2018 for citation, co-citation, co-publication and keyword co-occurrence networks. Authors suggested that citation network analysis effectively represents a measure of authors' influence and represents a path of scholarly communication. Co-citation analysis can identify influential researchers in a particular field of study. There is a strong "homology effect" in the co-citation network when two researcher nodes are very close, they are more likely to study on the same or similar research

topics, which facilitates research breakthroughs in the field. This research defined the collaboration network as a co-publication relationship between institutions and countries, which is an intuitive data of collaboration and is similar to a co-author relationship.

3.4.3 Other Collaborative Innovation Networks

This section discusses collaboration networks that cannot be presented in papers, such as co-development of systems, patent collaboration, email exchanges, and other special forms of collaboration.

The W3C (WWW Consortium) working groups consist of individuals and groups from one or more organisations, which then form the Collaborative Innovation Network (COIN) via the Internet. Gloor et al. (2003) [15] used the COIN mail analysis system architecture to query messages received or sent by a group during a certain time period, stored in a SQL database in a decomposed format, and then used the SNA visualization tool Pajek [5] and UCInet [8] to visualize the results of the SQL database queries.

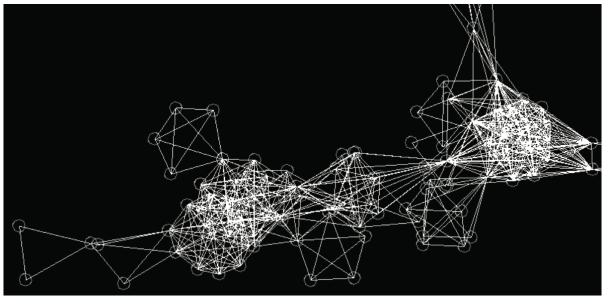
Prato et al. (2012) [13] used patent-based international cooperation data to study the dynamic evolution of technological research cooperation and develop a force-directed model of international technological cooperation. The authors suggest that the use of patents as a measure of international cooperation has limitations and therefore combine patent submission information such as inventor's personal information to construct a global technology cooperation network where the set of nodes is identified as countries and the bilateral relationship is calculated by the level of the national technology and the strength of the cooperative relationship to define the distance between two nodes.

Jermakovics et al. (2013)[19] conducted an analysis for software developer collaboration networks to calculate developer collaboration based on the modification records of common documents during open source software development. The cosine similarity between two developers illustrates that both share and modify the same files, and if there are no shared files then the similarity is 0. As collaboration information for open source projects is difficult to document, the automatic discovery of collaboration networks was developed in this study to make developer collaborations for open source projects clearer.

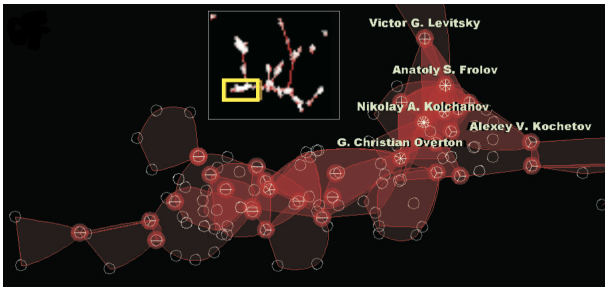
3.5 Visualization

In the visualization of clustered relationship mappings, we need to lay out nodes and establish connections between them, so the choice of layout algorithm is very important for the readability of large networks. There are five main node layout algorithms: minimum cut, smallest space (force-directed to establish attractive and repulsive forces between nodes), spectral/eigenvalue decommissions, tree/hierarchical and shape-based(Salter-Townshend et al., 2012) [36]. Force-directed mapping can be used to visualize the strength of relationships between researchers. This information can be used to identify potential collaborators, and prioritize research initiatives. In addition, networks with interactivity are useful for overlap visualization, keeping coarse granularity when there is no interaction, i.e. merging nodes that are too close to each other in the same cluster, and displaying a finer layout when interacting, where the user can effectively observe changes in granularity of different nodes when zoomed in/out and selected. There are many R packages for the SNA layout methods above, such as iGraph [12], RSIena [33], and Pajek [5].

Due to the complexity of real social groups, there is overlap between clusters and nodes, and when there are many nodes in a network graph and the size is large, the graph can be complex and difficult to find important information. Santamaría and Therón (2008) [38] therefore created a visualisation tool for overlapping that exploits variations in force-directed clustered graphs (FDCGs). FDCGs combines the spring forces formed by the internal and external edges of the clusters in the clustering diagram to generate the different gravitational forces between the nodes. Overlapper as a social network visualisation and analysis tool covers different graph analysis tools such as node-link (NL) diagrams, scatter diagrams, tree diagrams, etc. The focus of Overlapper is classification technique to determine whether there exists overlap between groups or not, which is shown in Figure 2. In addition, this paper also mentioned some popular tools which focus on force directed graph, such as GraphViz [14], Prefuse [17], and the statistical and mapping tools, e.g. Pajek [5] or JUNG [16].



(a) Traditional Node-Links representation



(b) Group drawing with Overlapper

Figure 2: The action of the overlapper in the same network. The edges of the graph are hidden and replaced by transparent hulls that wrap around the elements of each group. [38]

In 2014, Bian ey al. created CollaborationViz [6], which is an open source visual analytic tool for social network analysis (SNA) and research collaboration networks (RCN) in bio-medicine. CollaborationViz uses a web-based scalable vector graphics development framework to rank the influence of nodes in research collaboration networks, modelling influence in social networks using SNA centrality measures.

Collaboration Map (CoMap) (Lim & Chiu, 2015) [26] is an interactive collaboration network visualization tool that allows adjust the timeline to observe the collaboration records and cluster movement changes of researcher nodes over time. Nodes in CoMap are divided into two main types, locations and researchers. First CoMap aligns the location data of research publication to ensure the relative positions of different locations, and then renders the researcher nodes. When there is co-authorship between researchers, the relationship edges between the nodes will appear, and the gravitational force of the edges will be calculated based on the relationship parameters.

Venturini et al. (2021) [41] analyzed the ambiguity of social networks and presented a complete example of visual network analysis on jazz stakeholders, taking into account topological features in clustered layouts and discussing the advantages of preserving ambiguity in force-directed layouts. This paper delves into the interplay between relational ambiguity, and force-directed layouts in network analysis, and how visualizations can be used to effectively communicate insights about a network's structure and relationships.

4 Results and Discussion

Social Network Analysis (SNA) is a powerful tool for studying research collaboration and the relationships between individuals and organizations within a research community. By using SNA, researchers can gain insights into the structure of the collaboration network, the distribution of power and influence, and the dynamics of information flow within the network. SNA is a classical problem with extensive research and many theoretical and experimental foundations, but there is a lack of unified standards and algorithms for studying collaboration networks. This article provides a review of a significant number of pertinent studies for researcher publishing collaborations, primarily divided into co-authorship networks (direct relationships) and citation networks (indirect collaborations).

Co-authorship and citation networks can provide a platform to share information and resources, increase the efficiency and productivity of research, and help researchers identify key players and influential voices in their field.

The information gathered from the literature review process is used in this section to categorize and evaluate the software and algorithms employed in earlier studies on collaboration networks, as well as to discuss and address the research questions posed in section 2.2.

Table 3: Summary of the collaboration networks in this article

Author	Year	Data scope	Network	Description	Tool
Otte et al.	2002	Sociology	Co-authorship	- Information science-based SNA co-authorship networks and centrality metrics.	UCINET [8], Pajek [5]
Newman	2004	Biology, Physics, Mathematics	Co-authorship	- Proposed the calculation of clustering coefficients in co-authorship networks.	
Kajikawa et al.	2007	Sustainability	Citation	- Clustering analysis of sustainability based on citation networks.	LGL [3], VSM [37]
Santamaría et al.	2008	Computer Science	Co-authorship	- Presented Overlapper, a co-authored network visualisation overlapping group tool.	Overlapper [38]
Abbasi et al.	2011	Multi	Co-authorship	- Correlation between researcher position in co-authorship networks and research performance	AcaSoNet [1], UCINET [8]
Sun et al.	2011	Computer Science	Co-authorship	- A co-authorship prediction model PathPredic is proposed.	PathPredic [40]
Lim et al.	2015	Computer Science	Co-authorship	- A time-adjustable interactive co-authorship network visualization tool	CoMap [26]
Chen et al.	2016	InfoVis	Co-authorship	- Clustering of co-authorship networks using topic modelling	LDA [7]
Bu et al.	2016	Multi	Co-Citation	- Proposed the MACA framework based on author co-citation analysis (ACA).	MACA [9]
Alnajem et al.	2021	Circular Economy	Citation, Co-citation	- Provides a bibliometric analysis of citation, co-citation, collaboration and keyword co-occurrence networks in the CE field.	R version 4.0
Heldens et al.	2022	Multi	Co-authorship, Citation, Co-citation	- Python based research network analysis and processing tool	Litstudy [18]

RQ 1: What are the primary methods for obtaining accurate and up-to-date metadata for a collection of papers in the relevant literature? What are the most important metadata parameters to pay attention to?

- This paper summarizes the different data sources used by researchers in related studies, and compares the common bibliographic search APIs for base information and metadata whether they contain the data required for this experiment. In this study, we will perform co-authorship network analysis and citation network analysis for papers published by researchers in Dutch institutions, using topic modelling to cluster their research topics, with the main research topics obtained from keywords and abstracts in the papers. Therefore, the author, author ID, country, affiliation, keyword, abstract and citation information in the metadata are the key data for our study.

RQ 2: How to filter the data to study only researchers in the Netherlands?

- There are currently two ideas for a solution. The first is similar to the snowball sampling technique [21], which starts with the researchers at an institution (e.g. the eScience Center), classifies the research topics, and then finds more Dutch researchers (who are in the Dutch affiliations) in the same research area through the co-authors and citations of these studies. Then follow the trail of co-authors and repeat these procedures. However, it is obvious that this method is computationally intensive. The other approach is to get a series of literature with similar subjects, by searching keywords or topics. And then filter it to papers where at least one author is at a Dutch affiliation. Therefore, the metadata set for our study needs to contain at least the country or institution of the author.

RQ 3: What is the role of topic modelling in the overall workflow and what advanced technologies are currently available that could be useful for our research?

- A closed and dense social network of research should focus on the same research area. Topic modelling provides a method for clustering literature from different research areas, extracting research topics against keyword and abstract data and ensuring the centrality of the social network. Latent Dirichlet Allocation (LDA) is often applied in the topic modelling process in related studies.

RQ 4: How to define the strength of relationships between researchers in a force-directed graph?

- For researcher communication, we will focus primarily on co-authorship and citation relationships. When author A and author B jointly publish research, they have a co-authorship relationship and the strength of the relationship is related by the number of times they collaborate. The citation relationship, on the other hand, requires counting the frequency of citations and citations between author A and author B. Their co-citation information can also be taken into account.

RQ 5: What are some common tools for visualising social networks?

- There are a number of excellent tools for visualising social networks such as UCINET[8], Pajek[5], RSIena [33], which are also tools for complex network analysis. In addition, iGraph [12], GraphViz [14], Prefuse [17] etc. also offer network image visualisation techniques.

5 Conclusion

In conclusion, SNA is a valuable tool for understanding and optimizing research collaboration, by providing insights into the structure of the collaboration network, the distribution of power and influence, and the flow of information within the network. By using SNA, researchers can gain a deeper understanding of their collaborations, identify opportunities for improvement, and optimize their collaborations for maximum impact.

Geographically-focused social networks for academic papers are a relatively new field; the majority of current research on social networks for academic collaboration focuses on in-depth analyses of communication networks in a specific discipline or on thematic clustering analyses of a specific dataset. In addition to non-published research collaborations, social network analysis of research collaboration networks is subdivided into the analysis of paper co-authorship networks and the analysis of paper citation networks, where co-citation networks are also covered, for this literature review. This paper compares the suitability of currently popular literature search engine APIs to the subject of this study for the purpose of research data acquisition. Experiments will also be carried out to further refine the data selection in the subsequent research. Finally, there are many SNA solutions that have not been addressed because the data and analysis presented in this literature study fall far short of thoroughly exploring the enormous discipline of SNA. This paper only discusses some research that has been done so far in relation to the collaboration network of the researcher, with the goal of giving readers an overview of this particular field of study.

References

[1] Alireza Abbasi and Jorn Altmann. A Social Network System for Analyzing Publication Activities of Researchers. TEMEP Discussion Papers 201058, Seoul National University; Technology Management, Economics, and Policy Program (TEMEP), April 2010.

[2] 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.

[3] Alex T. Adai, Shailesh V. Date, Shannon Wieland, and Edward M. Marcotte. Lgl: Creating a map of protein function with an algorithm for visualizing very large biological networks. *Journal of Molecular Biology*, 340(1):179–190, 2004.

[4] 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.

[5] Vladimir Batagelj and Andrej Mrvar. Pajek-program for large network analysis. *Connect*, 21:47–57, 01 1998.

[6] 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.

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

[8] Stephen P. Borgatti, Martin G. Everett, and Linton C. Freeman. *UCINET*, pages 2261–2267. Springer New York, New York, NY, 2014.

[9] Yi Bu, Tian-yi Liu, and Win-bin Huang. Maca: a modified author co-citation analysis method combined with general descriptive metadata of citations. *Scientometrics*, 108, 05 2016.

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

[11] Francine Chen, Patrick Chiu, and Seongtaek Lim. Topic modeling of document metadata for visualizing collaborations over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, page 108–117, New York, NY, USA, 2016. Association for Computing Machinery.

[12] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[13] Giuditta De Prato and Daniel Nepelski. Global technological collaboration network. network analysis of international co-inventions. *J Technol Transf*, 39:1–18, 12 2012.

[14] J. Ellson, E.R. Gansner, E. Koutsofios, S.C. North, and G. Woodhull. Graphviz and dynagraph – static and dynamic graph drawing tools. In M. Junger and P. Mutzel, editors, *Graph Drawing Software*, Mathematics and Visualization, pages 127–148. Springer-Verlag, Berlin/Heidelberg, 2004.

[15] Peter Gloor, Robert Laubacher, Scott Dynes, and Yan Zhao. Visualization of communication patterns in collaborative innovation networks - analysis of some w3c working groups. page 56, 01 2003.

[16] Greg. *JUNG 2.0 Tutorial*, 2010.

[17] Jeffrey Heer, Stuart K. Card, and James a. Landay. Prefuse: a Toolkit for Interactive Information Visualization. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '05*, page 421, 2005.

[18] Stijn Heldens, Alessio Sclocco, Henk Dreuning, Ben van Werkhoven, Pieter Hijma, Jason Maassen, and Rob V. van Nieuwpoort. litstudy: A python package for literature reviews. *SoftwareX*, 20:101207, 2022.

[19] Andrejs Jermakovics, Alberto Sillitti, and Giancarlo Succi. Exploring collaboration networks in open-source projects. volume 404, pages 97–108, 06 2013.

[20] Yuya Kajikawa, Junko Ohno, Yoshiyuki Takeda, Katsumori Matsushima, and Hiroshi Komiyama. Creating an academic landscape of sustainability science: An analysis of the citation network. *Sustainability Science*, 2:221–231, 09 2007.

[21] C D Kaplan, D Korf, and C Sterk. Temporal and social contexts of heroin-using populations. an illustration of the snowball sampling technique. *J Nerv Ment Dis*, 175(9):566–574, September 1987.

[22] J.Sylvan Katz and Ben R. Martin. What is research collaboration? *Research Policy*, 26(1):1–18, 1997.

[23] Heiko Koziolek. *Goal, Question, Metric*, pages 39–42. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

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

[25] Kar Wai Lim and Wray Buntine. Bibliographic analysis with the citation network topic model. 2016.

[26] Seongtaek Lim and Patrick Chiu. Collaboration map: Visualizing temporal dynamics of small group collaboration. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work amp; Social Computing*, CSCW'15 Companion, page 41–44, New York, NY, USA, 2015. Association for Computing Machinery.

[27] Peter V. Marsden. *Survey Methods for Network Data*, pages 370–388. Sage Publications, London, 2011.

[28] Newman and Mark J. Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101:5200–52005, 01 2004.

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

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

[31] M. E. J Newman. Fast algorithm for detecting community structure in networks. phys. rev. e stat. nonlin. soft. matter. phys. 69(6 pt 2), 066133. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69:066133, 07 2004.

[32] 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.

[33] Ruth M. Ripley, Tom A. B. Snijders, Zsofia B'oda, Andr'as V"or"os, and Paulina Preciado. Manual for Siena version 4.0. Technical report, Oxford: University of Oxford, Department of Statistics; Nuffield College, 2023. R package version 1.3.14.1. https://www.cran.r-project.org/web/packages/RSiena/.

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

[35] Brendan Rousseau and Ronald Rousseau. Lotka: A program to fit a power law distribution to observed frequency data. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics, ISSN 1137-5019, Nº. 4, 2000*, 4, 01 2000.

[36] 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.

[37] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, nov 1975.

[38] Rodrigo Santamaría and Roberto Therón. Overlapping clustered graphs: Co-authorship networks visualization. In *Proceedings of the 9th International Symposium on Smart Graphics*, SG '08, page 190–199, Berlin, Heidelberg, 2008. Springer-Verlag.

[39] 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.

[40] 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] Tommaso Venturini, Mathieu Jacomy, and Pablo Jensen. What do we see when we look at networks: Visual network analysis, relational ambiguity, and force-directed layouts. *Big Data & Society*, 8(1):20539517211018488, 2021.