PhD-FSTM-2021-036
The Faculty of Sciences, Technology and Medicine

DISSERTATION

Defence held on 17/06/2021 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

en *Informatique*

by

Antonio Maria FISCARELLI
Born on 18 May 1990 in Foggia (Italy)

SOCIAL NETWORK ANALYSIS FOR DIGITAL HUMANITIES

Dissertation defence committee

Dr Pascal Bouvry, dissertation supervisor
Professor, Université du Luxembourg

Dr Grégoire Danoy
Université du Luxembourg

Dr Andreas Fickers, Chairman
Professor, Université du Luxembourg

Dr Christophe Verbruggen
Professor, University of Gent

Dr Thomas Stützle, Vice Chairman
Professor, Université libre de Bruxelles

ABSTRACT

Fiscarelli, Antonio Maria (Ph.D., Computer Science)

Social network analysis for digital humanities

Dissertation directed by Professor Pascal Bouvry

Current trends in academia show that a key factor for tackling complex problems and doing successful research is interdisciplinarity. With the increasing availability of digital tools and online databases, many disciplines in the humanities and social sciences are seeking to incorporate computational techniques in their research workflow. Digital humanities (DH) is a collaborative and interdisciplinary area of research that bridges computing and the humanities disciplines, bringing digital tools to humanities scholars to use, together with a critical understanding of such tools. Social network analysis is one of such tools. Social network analysis focuses on relationships among social actors and it is an important addition to standard social and behavioral research, which is primarily concerned with attributes of the social units.

In this work we present the field of digital humanities and its current challenges, as well as an overview of the most recent trends in historical network research, emphasizing the advantages of using social network analysis in history and the missed opportunities. We then present the field of network analysis, providing a formalization of the concept of social network, models that explain the mechanism governing complex networks and tools such as network metrics, orbit analysis and Exponential Random Graph Model.

We tackle the problem of community detection. We propose MemLPA, a new version of the label propagation algorithm, by incorporating a memory element, in order for nodes to consider past states of the network in their decision rule. We present a use case, drawn from the collaboration with a historian colleague, showing how social network analysis can be used to answer research questions in history. In particular, we addressed the gender and ethnic bias problem in computer science research by looking at different collaboration patterns in the temporal co-authorship network. Finally, we present another use case, based on collaboration data collected at the National Electronics and Computer Technology Center (NECTEC) in Thailand. We build a temporal collaboration network where researchers are connected if they worked together on one or more artifacts, focusing on measuring productivity and quality of research and development, while linking these metrics to the structure of the collaboration network.

“For the first time in his life, Grakk felt a little warm and fuzzy inside.”

ACKNOWLEDGMENTS

I would like to thank my supervisor, Prof. Pascal Bouvry, for his trust and confidence in my work. I would like to thank Dr. Matthias Brust, who has been not only a mentor but a dear friend. I want to thank you for your patience and for supporting my ideas. I would also like to thank my colleagues Sytze, Lorella, Sea and Roland for their wonderful collaborations. Our conversations have been very inspirational and you helped me explore new areas of research. I would like to express my gratitude to the Luxembourg National Research Fund and the Luxembourg Centre for Contemporary and Digital History (C2DH) for financing this project.

I would like to thank my parents, sister and extended family. I highly regard them for their unconditional love ans support in everything I do. I would also like to thank Valeria, for providing all the love and care that was required for me to stay on tracks. Finally, I would like to thank Gianluca, Andreina, Alice, Giorgio and all my friends. In particular, I would like to single out Tom, my best friend in Luxembourg, whose support was always there when needed.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

CHAPTER I

INTRODUCTION

Networks are all around us. Many complex systems such as the Internet, biological systems and transportation systems can be modelled using networks. These so-called complex networks exhibit properties that are not found in “artificial networks”. For instance, information flows faster in a complex network when compared to a random network. Complex networks are particularly resilient to random failure of nodes, while at the same time are more sensitive to targeted attacks. Finally, complex networks often exhibit a community structure, where nodes in the same community are highly connected to each other and loosely connected to the rest of the network. From a structural point of view, all these properties translate into networks having an unequal degree distribution where few nodes are highly connected and the majority has few connections only, short distances between nodes and high level of transitivity.

Social relationships can also be represented using networks. Social network platforms such as Facebook and Instagram are the most common examples, where people connect to each other based on friendship or common interests. Social networks can be used to model the dynamics of collaboration between scientists in academia or professionals in the private sector. These special networks are called collaboration networks or co-authorship networks. Nodes represent individuals that are connected to each other if they worked on the same project or co-authored a scientific article. The field of Science of Science is a new emerging field whose aim is to understand the circumstances that lead to successful research. The most common metrics to measure individuals' performance, such as the h-index, are based on number of publications and number of citations. These metrics may not be sufficient to provide a deep insight into the factors driving scientific success. Furthermore, the different authorship and citation practices in each field make these metrics unsuitable and not universally valid. Instead, network analysis techniques based on centrality measures could be used to shed a new light on some mechanisms of success. On the other hand, in recent years, interest moved from individuals to teams. It has become more and more apparent that the most successful research is carried out by teams rather than single researchers. For this reason, researchers have become interested in studying the dynamics of social groups, and new performance measures are necessary to quantify teams' success.

1.1 Motivation

This project is collocated within the Doctoral training unit (DTU) in digital history and hermeneutics [1]. The main objective of the DTU is to create a trading zone for the reflection of epistemological and methodological challenges in digital history, where scholars from humanities disciplines such as History, Philosophy, Psychology, Linguistics and Archaelology, as well as computer scientists, find a common ground, shape a common language and negotiate new forms of knowledge in an interdisciplinary setting. One of the objectives of this work is to show how humanities research can benefit from network analysis, by providing Ph.D. students from other disciplines with the right tools that can help them answer their historical questions and adapting these tools to their research projects. In this way, we seek for a fruitful collaboration where both sides can benefit from each other: humanities scholars gain a critical understanding of digital tools and their functionalities, while computer scientists find new use cases and applications while learning to understand the needs of humanists. Understanding each other's needs is crucial for the collaboration. Instead of two distinct figures with separate interests, I envision humanists and computer scientists join forces and share their knowledge and expertise in order to tackle the new challenges that are emerging in digital humanities. Only with a common goal and a shared vision can this collaboration be effective and still worth the effort and time required.

1.2 Research contributions

This work makes the following contributions:

- Introduces the field of digital humanities and its current challenges, as well as an overview of the most recent trends in historical network research, emphasizing the advantages of using social network analysis in history and the missed opportunities. It then present the field of network analysis, providing a formalization of the concept of social networks, models that explain the mechanism governing complex networks and tools such as network metrics, orbit analysis and Exponential Random Graph Model.
- Introduces MemLPA, a new version of the label propagation algorithm, that incorporates a memory element, in order for nodes to consider past states of the network in their decision rule. The algorithm was tested on both artificial and real world networks, using classical performance metrics, as well as metrics to quantify the structural characteristics of the communities found. MemLPA outperforms all other label propagation algorithms that implement a memory mechanism, as well as some of the more complex state-of-the-art community detection algorithms. This is achieved while being completely scalable, using local interaction only and running in linear time.

- Presents a use case, drawn from the collaboration with a historian colleague, showing how social network analysis can be used to answer research questions in history. In particular, we addressed the gender and ethnic bias in computer science research by looking at different collaboration patterns in the temporal co-authorship network. We started with the following research questions: “Do men and women, as well as researchers of different ethnies, show differences in collaboration patters? How do these differences shape the network of collaboration?” We answered these questions by using network metrics that are based on researcher’s position in the network and their neighborhood structure, rather than classical performance metrics based on number of publications and number of citations. We found that the women score lower than men in terms of performance metrics and are more close-knit. Women are shown to occupy peripheral positions in the network, while men are more likely to cover central positions. Researchers of color score higher in ranks, while white researchers are more close-knit. Looking at the evolution of the co-authorship network, we showed that differences in gender and ethnicity are narrowing over time.
- Presents a use case based on collaboration data collected at the National Electronics and Computer Technology Center (NECTEC) in Thailand. Researchers collaborate on different projects and team up to produce a range of artifacts. For each artifact, a score that measures quality of research is available and shared between the researchers that contributed to its creation, according to their percentage of contribution. We build a temporal collaboration network where researchers are connected if they worked together on one or more artifacts. We started with the following research questions: ”What are the collaborations patterns that lead to individual and group success?” We answered this question by measuring productivity and quality of research and development, while linking these metrics to the structure of the collaboration network. We found that researchers that cover more central positions in

the network are more performing. At the same time, centrality metrics are not found to be correlated with average IC score, which measure quality of work rather than quantity. For what concerns teams, we proposed some team structural metrics that can be used to assess team's performance. In particular, we found that team density, turnover and openness are positively linked to team performance.

1.3 Publications

1. Fiscarelli, A. M., Brust, M. R., Danoy, G., & Bouvry, P. (2018). *A memory-based label propagation algorithm for community detection. In International Conference on Complex Networks and their Applications* (pp. 171-182). Springer, Cham.
2. Fiscarelli, A. M., Brust, M. R., Danoy, G., Bouvry, P. (2019). *Local memory boosts label propagation for community detection. Applied Network Science*, 4(1), 1-17
3. Fiscarelli, & Van Herck, S. (2018). *Minorities in computer science. Gender and ethnic collaboration patterns in a temporal co-authorship network. Submitted to PloS one*
4. Fiscarelli, A. M., Brust, M. R., Bouffanais, R., Piyatumrong, A., Danoy, G., & Bouvry, P. (2021). *Interplay between success and patterns of human collaboration: case study of a Thai Research Institute. Scientific Reports*, 11(1), 1-14.
5. Antonio M. Fiscarelli, *Social network analysis for digital humanities: challenges and a use case.* In Andreas Fickers, Juliane Tatarinov (eds.), Digital History and Hermeneutics: Theory and Practice. Berlin: De Gruyter 2021 (forthcoming)

CHAPTER II

SOCIAL NETWORK ANALYSIS FOR DIGITAL HUMANITIES

This chapter is partially based on the article “Social network analysis for digital humanities: challenges and use cases” [2], as part of the book “Digital History and Hermeneutics” [3].

Digital humanities (DH) is a collaborative and interdisciplinary area of research that bridges computing and the humanities disciplines, bringing digital tools to humanities scholars to use, together with a critical understanding of such tools. Thanks to the availability of online digital sources, new software and tools, the field of Digital Humanities has been growing in the last decades. Still, being a novel area of research, there are some challenges that must be faced.

At the same time, due to the growing computing power and availability of online databases, network analysis has gained popularity among researchers from different fields, who have jumped on the network science “wagon” and words such “network” and “complexity” have become more and more commonly used.

A social network can be described as a collection of social actors that are connected to each other if they form some sort of relationship. Social network analysis focuses on relationships among social actors and is an important addition to standard social and behavioral research, which is primarily concerned with attributes of the social units [4]. Not only is it important to acknowledge that social relationships are relevant, but also to understand how these ties work and how they relate to the many underlying social mechanisms governing these networks.

Social network analysis belongs to the tools that have become very popular amongst humanities scholars. Even though social networks are thought to be a modern invention, due to the popularity Facebook and other online platforms, they are not limited to the modern days [5]. For example, social networks have been used to model marriage and business relationships of the Medici family in the 15th century Florence [6], the evolution of women's social movements in the 19th century [7], the personal support network of Jewish refugees during the second world war [8] and visibility networks of long barrows in Cranborne Chase [9].

2.1 Challenges in digital humanities

2.1.1 Drawing complicated graphs

The first challenge in Digital Humanities is of methodological nature [10]. On the one hand, especially for the use of network analysis, there is a risk that humanities research will limit itself to the "drawing of complicated graphs" [11]. A certain method or the use of a digital tool can not be the main objective of research. On the other hand, some scholars may be hesitant to introduce digital tools in their research, fearing that this will take them out of the history realm. Therefore, it is important to understand what digital tools can really offer to support historical research.

2.1.2 Black boxes and data providers

The second one is related to the interdisciplinary nature of digital humanities. Humanities research can manifest in two forms. In one case, scholars may show interest in a digital tool, start experimenting with it and include it in their workflow. This approach could lead to the use of such tools as a “black box” [12]. Given some input, this black box will produce a certain output while everything in between is unknown. Therefore, it will not be possible to understand how a certain tool works, how to interpret the results and recognize any possible bias that a tool inherently has. In the second case, scholars may seek for help or collaboration with an expert, a computational expert with a solid background on a specific method or tool. In this case, there is the risk for the scholar to become a simple “data provider” for the model maker [11].

Even when the scholar uses a tool, there is an indirect interaction between them and the computational expert. In fact, the choices made by the developers of such tool, and its functionalities, directly affect the user experience. I argue that a direct interaction between the two parties, who can actively engage in discussions and experiment actively, is more beneficial to the collaboration.

It is also essential to find a common vocabulary and be able to conciliate the two different perspectives. Only if this is achieved, the two researchers can start negotiating new forms of knowledge and successfully do historical research.

2.1.3 Data availability

Another issue is related to the data itself. Historians nowadays have access to much larger amount of data, coming from digitized classical sources (book scans, digitized old photographs, recordings) and digitally born sources (websites, social networks). Not only do they have have access to large amount of data, they can access it at high speed and relatively low cost. For that reason, historians may be experiencing a fundamental paradigm shift, going from a scarcity of sources to an abundance of sources [13], while traditional methods used by historians may fail to deal with such amount of information. One example of such methods is close reading, that may fail its purpose for very large collections of texts without the support of computer-based techniques. Easy accessibility of data comes with new questions too. What sources have been digitized, what were discarded and what criterion was used to select them? It is also important to identify the origin of such sources. What was the provenance of the original sources? For what concerns digitally born source, how have they been generated?

2.1.4 Data storage and accessibility

Data storage has also changed with the advent of the digital era. The use of new technologies has made storing data much easier. A single hard drive can store thousands of documents, is cheap, small and easy to transport. One may think that digital data can last forever. Unfortunately, data stored in a digital form does not have any intrinsic meaning without a specific software or technology that can read it, and these technologies can become obsolete in a decade or even less. One may also think that digitally stored data is safe from aging. Unlike analog sources, digital data does not deteriorate. On the other hand, a single malfunction of the storing volume could make the entire collection inaccessible and lost forever [14].

2.2 Current trends in historical network analysis

There are several examples of historians incorporating network analysis in their research, and the number of journals dedicated to digital humanities is an indicator of the growing field. One example is the Journal of Historical Network Research, a fully Open Access journal focused on networks and network research in history, published in cooperation with the Luxembourgish Centre for Contemporary and Digital History (`https://www.c2dh.uni.lu/`) and `historicalnetworkresearch.org`. It collects papers centred on historical networks of any period of the recorded human past, from Bronze Age civilisation to contemporary history. For this section, I picked up some articles that I believe are worth looking at, as they are good examples for anyone interested in knowing how historical research can be done using network analysis tools. I review these works, I show how they translated historical questions into a social network analysis perspective, and I identify the missed opportunities in the studies.

2.2.1 Reconstructing science networks from the past

Breure and Heiberger, in their study [15], argue that eponyms serve as a proxy for contact and are a promising way to explore historic relationships between natural scientists. Eponyms are used in taxonomy when an author describes new species for which he uses the name of a person. Eponyms are normally given to field collectors or colleagues. They tested this hypothesis on the community of malacologists in the 19th century, analysing the record activity of malacological authors between 1850 and 1870.

The dataset used contains authors' information such as age and home country, as well as performance measures such as number of publications, number of pages, number of co-authored publications and number of co-authors. Each connection between authors is classified as eponyms, exchange of material or co-authorship. Therefore, the authors built a collaboration network, in particular a multiplex network, where nodes interact in different layers (depending on the type of interaction) but there is no interaction between different layers. The network, consisting of 476 nodes and 1,822 edges, can be considered of medium size.

Authors in the network are ranked according to their number of publications, and elite authors are identified as the authors who contributed to 80% of the total publications. Breure and Heiberger noticed that few authors publish a large number of papers, something that has been widely recognized in bibliometrics as the skeweness of science [16], although they did not show the actual distribution of authors' productivity. The authors recognized two densely connected communities that represent recent and palaeontological authors. They manually assigned authors to one of the two communities, depending on their research interests. It would have been interesting to use a community detection algorithm and compare the communities found with the ones identified by the authors, using metrics such as normalized mutual information [17, 18] or adjusted randomized index [19] to quantify the agreement of the result, in order to assess any bias in the manual assignment. The authors used Exponential random graph models (ERGMs) [20–22] to find out what effects shaped the network of collaboration. They found that authors from the same country are more likely to connect and that publications increase the odds of a tie between authors. They also discuss how eponyms could result in a collaboration between authors. This hypothesis was not tested, even thought ERGMs gives the possibility to test whether a tie in one layer of the network increases the odds of a tie in a different layer.

2.2.2 Network analysis of medieval manuscript transmission

Fernandez, in his work [23], introduces a new method, based on network analysis, to analyse shared manuscript transmission of medieval German texts. Medieval manuscripts contained several texts, that were brought together according to certain criteria, both cultural (common genre) and practical (availability, size, etc), rather than being randomly grouped together.

The author modelled the transmission of shared manuscripts as a network, where nodes represent texts that are connected if they appear in the same manuscript, and a weight is used if texts appear together in more than one manuscript. The author does not mention the size of the network, however he specifies that the giant component, the largest connected component of the network, includes 76% of the nodes, while several smaller components (two to eight nodes) include 6% of the nodes and the remaining 18% of the nodes consists of isolated nodes. He decided to name these three different parts of the network “Continent”, “Archipelagos” and “Islands”.

The author proceeded by applying a community detection algorithm (see section 4.1) on the largest component to identify communities, although the algorithm used is not mentioned. Since nodes have no attribute data available, such as genre, time or location, the author inspected manually the outcome of the algorithm to verify whether any of these characteristics correlates with the communities found, and came to the conclusion that there is a high overlap between communities, even for different genres. He used Eigenvector centrality to identify texts that tend to appear in big collections and betweenness centrality to identify texts that connect different communities in the network and fit in different genres. These metrics helped the author identify texts that cover important positions in the network, something that would have been impossible with a human inspection (see section 3.2.4 for an overview on centrality metrics).

In the end, even though the author does not really provide statistical methods to analyse the network of interest as he mentioned, limiting his work to the visualization of the network and the computation of centrality metrics, it must be recognized that the data available was rather limited.

2.2.3 The emergence of epistemic communities in the “Sphaera” corpus: mechanisms of knowledge evolution

Valeriani et al. [24] analyzed the emergence of epistemic communities during the early modern period. They worked on a corpus of printed cosmology textbooks used in European universities. Each book was divided into several text parts, representing atoms of knowledge.

They built a directed, weighted, multi-layer network where nodes represent books that are connected to each other, on different layers, if they contain text parts that re-occur in time (they contain the same text, adaptations of the same text, translations of the same text, commentaries of the same text, commentaries of the same adaptation), for a total of five layers. The network is directed, with direction being chronological, from older to more recent occurrence. The weight of connections, instead, is given by the number of text parts that re-occur in two different books. The corpus contains 563 text parts, but the authors decided to consider only the ones that reoccur at least once and at least one year between re-occurrence. Therefore the network, which can be considered of small-medium size, consists of 239 text parts ad 1,625 re-occurrences. The authors also analyse the aggregated graph, which includes the same set of nodes and two nodes are connected if they are connected in any of the five layers.

The authors perform a longitudinal analysis, by first looking at the age distribution of connections for each layer of the network, computed as the difference between year of publication of the two text parts at the ends of each connection, finding substantial difference between layers. Then, they look at the different connected components of the network in order to identify the different epistemic communities. With a series of plots, they analyze the distribution of nodes' out-degree, normalized by the publication time of the text. For each plot, the visualization is further enhanced with different colors representing nodes' attributes such as in-degree, publication place, book format and network layer. The analysis is followed by an in depth interpretation of the results and discussion on the emergence and evolution of the different families of editions.

Again, the methodology provided is based more on data visualization rather than statistical analysis.

2.2.4 Athens as a small world

Cline, in her work [25], uses social network analysis to study the political life in Athens between the 460s and 450s BC.

She built three different social networks using Plutarch's "Lives", from which she retrieves all actors and how these are related to each other. The first one uses Plutarch's

Life of Pericles and consists of 54 actors and 79 ties, which is basically Plutarch's ego network. She then enlarges it by including Plutarch's Life of Alcibiades. The second version of Athens' social network contains 106 nodes and 145 connections. Finally, she includes Plutarch's Lives of Cimon and Nicias, for a total of 133 nodes and 191 ties. These networks are all of small size, undirected and unweighted. The author claims to be working with a multiplex network, since ties between actors are of different nature (family, work, friendship), even though there is no distinction between these ties in the analysis.

The objective of the author is to demonstrate that the social network of Athens' political life is a small world (see section 3.3.2). Her argument is that democratic institutions in Athens enabled people belonging to different circles and social classes to meet, hence favouring innovation and the diffusion of new ideas. From a network perspective, this would reflect in Athens' social network having a low average path length, high level of transitivity and a core-periphery structure were degree distribution follows a power law, with few highly connected nodes and most nodes with low degree. Indeed, she computes transitivity, average path length and diameter for all the networks, and compare them with the same quantities computed on a random network having the same size. All these quantities confirm that Athens is a small world. For what concerns the core-periphery structure, she computes the degree distribution but does not perform any statistical test to verify whether a power law is the best fit. The author also computes betweenness centrality (see section 3.2.4) for each actor to confirm that women tend to cover central positions in the network, connecting different families thanks to arranged marriages. For this work, information such as gender, family and social status was available but not analyzed. It would have been interesting, for example, to test the level of homophily in the network, i.e. whether two nodes sharing the same gender/family/status increased the likelihood of a connection between them.

2.2.5 Searching for hidden bridges in co-occurrence networks from Japanese Wayang Kulit

Schauf and Varela [26] use network analysis techniques to identity characters that covered structural roles in the Japanese Wayang Kulit incarnation of the Mahabharata epic. The Mahabharata consists of a series of stories and their representations, called lakon, from the epic.

They build a weighted, undirected co-occurrence network, where nodes represent the characters of the epic and these characters are connected if they are mentioned in the same scene. Weights indicate how many times two characters appear in the same scene. Each node is enriched with several attributes such as characters' tribe affiliation, origin, species and gender. The authors also build two different null models that preserve, on average, the degree distribution of nodes.

They compute betweenness centrality and closeness centrality (see section 3.2.4) for each character in the empirical network, as well as in the two null models. In this way, it is possible to identify outliers whose centrality values are significantly higher or lower than expected, i.e compared to the same quantity computed in the null models. They find that female characters, despite being few in number and appearing relatively infrequently, appear to dominate the top ranks for betweenness. They also propose a variation of these centrality metrics that is based on nodes' attributes. For example, the inter-faction betweenness centrality is used to identify those characters who act as bridges withing their tribe, while the faction-world betweenness centrality identifies characters who act as bridges between their tribe and the rest of the network.

2.3 Missing data in historical network research

One of the challenges that emerges by analyzing the current trends in historical network research is missing data and incomplete data [5]. Networked data has to be extracted from sources such as books, bibliographies and diaries that are analogic and only digitized afterwards if needed. These sources are often incomplete or does not provide enough information to build the network of interest. Also, missing data in network research is more critical when compared to social and behavioral research. Even a small portion of missing data can be problematic, if this data is related to crucial nodes (see hubs in Section 3.3.3) or ties (see weak ties in Section 3.4). This also goes in contrast to historical research working with digitally born data, such as online databases or data scraped from social networks, where data is rather abundant.

2.4 Doctoral training unit in Digital History and Hermeneutics

This project is part of the Doctoral training unit (DTU) in Digital History and Hermeneutics (`https://dhh.uni.lu/`) [1]. The objective of the DTU is to create a trading zone for the reflection of epistemological and methodological challenges in digital humanities, where scholars from different disciplines find a common ground, shape a common language and negotiate new forms of knowledge in an interdisciplinary setting.

Trading zones are either physical or virtual spaces where two communities with different practices and inter-languages, in this cases the humanities and the computer science community, can interact. [27]. These trading zones allow to cross the disciplinary boundary of a community, without losing their own identity and practices. For instance, historians can share local understandings of concepts from computer science, without needing to understand the entire complexity of computer science, or become computer scientists themselves [28]. The same applies for computer scientists, that do not need to know the common practices in history to tackle problems in history applications.

2.4.1 Use cases

This work includes two use cases, with the purpose of showing how social network analysis can be used to answer research questions in Digital Humanities. Chapter VI presents a use case, drawn from the collaboration with the National Electronics and Computer Technology Center (NECTEC) in Thailand, where I investigate the collaborations patterns that lead to individual and team success in the collaboration network of the institute.

Chapter V presents a use case drawn from the collaboration with a historian colleague. Sytze Van Herck is one of the Ph.D. students of the Doctoral training unit in Digital History and Hermeneutics. Her main research interests are intersectionality and gender within the history of computing [29, 30]. Her work examines occupational segregation, working conditions and gender stereotypes in advertising from the 1930s until the end of the 1980s [31–33]. Together with Sytze we have applied social network analysis techniques to analyse the gender and ethnicity gap in the Computer Science research community [34]. In particular, I present our research questions and the tools that we have used to answer them. Finally, I present a reflection on the challenges that Sytze and I encountered during our joint work such as the generalizations that we made to model our scenario and the algorithm criticism regarding the gender/ethnicity prediction.

CHAPTER III

COMPLEX NETWORKS

Social networks [4], the world wide web [35], biological networks [36] and many other real world networks show properties that are not found in “artificial” networks. For example, information flows faster in complex networks, they are particularly resilient to random failure of nodes, [37], and they often exhibit a community structure. This is a consequence to the fact that artificial networks are designed by humans, who have a centralized control over the building process of the network. Many real-world networks, instead, are born spontaneously from the local interaction between entities. Some of these special properties played an important role in the development of the field and will be presented in this chapter.

3.1 Graphs

Historically, the first encounter with graph theory refers to the so called “Seven bridges of Königsberg” problem [38]. The city of Königsberg was built on four main areas: the two sides of the Pregel River and two small islands, connected by seven bridges. The problem consisted in finding a path that reached all the areas of the city by crossing each bridge exactly once. Euler modelled this problem using graph theory, representing city areas as nodes and bridges as edges connecting nodes, and proved this problem to be unfeasible.

More formally, a graph $G = (V, E)$ is a pair of sets where V represents the set of vertices, or nodes, and E represents the set of edges, or links, that connect nodes. Graphs can also be represented using a adjacency matrix A, where element a_{ij} is different from zero if there exists and edge e_{ij} connecting nodes v_i and v_j. A graph can be directed or undirected, whether the direction of a connection is relevant or not. In case of an directed graph, elements of the associate adjacency matrix will be positive or negative depending on the direction of the connection, or just positive for undirected graphs. A graph can also be weighted or unweighted, where weight represents cost, strength or importance of a connection. In case of a weighted graph, the elements of the adjacency matrix can assume any value, while they can only assume zero or one for an unweighted graph. Notice that the terms network and graphs are often used interchangeably.

3.2 Network metrics

3.2.1 Degree

The *degree* of a node v_i represents the number of incident edges of a node, or equivalently the number of non-null elements of i-th row of the adjacency matrix as showed in equation III.6. In other words, it represents the number of direct connections of a node. In case of directed network, in-degree and out-degree are defined, and they refer to the number of in-going or out-going edges of a node. In a social network, the degree of a node represent the number of direct connections, i.e. the number of neighbors, friends or followers.

$$Deg(v_{ij}) = \sum_{j=1}^{N} a_{ij} \tag{III.1}$$

3.2.2 Average path length and diameter

The *average path length* of a network is defined as the average shortest path between any two nodes in a network. Given a matrix S, where the element s_{ij} represents the length of the shortest path from node v_i to node v_j, the average shortest path of a network is computed as the average value of each element of matrix S as showed in equation III.2.

$$APL(G) = \frac{1}{N(N-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} s_{ij} \tag{III.2}$$

The *diameter* of a network, instead, is defined as the maximum shortest path in a network. It is computed as the maximum element of matrix S as showed in equation III.4. In a social network, for example. these two metrics represent how easily information/news/ideas spread in the network.

$$D(G) = \max_{i,j \in 1..n} (s_{ij}) \tag{III.3}$$

3.2.3 Clustering coefficient

The *clustering coefficient* of a network is defined as the average local clustering coefficient of each node in the network. The local transitivity of a node is the ratio of the triangles connected to the node and the triples centered on the node [39]. This metric is related to the concept of transitivity: given that v_i is connected to v_j and v_j is connected to v_k, what are the odds that v_i is also connected to v_k? In a social network, for example, transitivity measures the degree to which the friend of a friend is also your friend. Social networks, particularly, show high transitivity, when compared to a random network.

$$C(G) = \frac{1}{n} \sum_{i=1}^{n} c_i = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{k_i(k_i-1)} \sum_{j,h}^{n} a_{ij} a_{ih} a_{jh} \right) \tag{III.4}$$

3.2.4 Centrality metrics

Centrality metrics are an important tool for the analysis of social networks. They are defined on the nodes and they rank nodes according to their position in the network [40]. The first person to experiment with centrality metrics was Bavelas [41, 42], who showed that centrality measures were linked to group performance and that centrality metrics can help identify people with different roles in the network.

3.2.4.1 Degree centrality

Degree centrality correspond to the degree on a node. It measures the number of direct connections of a node and can be used to identify actors that are highly connected.

3.2.4.2 Betweenness centrality

Betweenness centrality is computed as the number of shortest paths between any two nodes in the network that go through a certain node. It measures to what extent an actor has control over the information flowing between others and can be used to identify actors who occupies strategic positions in the network in terms of information exchange. Betweeness centrality is computed as following:

$$Bet(v_i) = \sum_{j \leq k} \frac{g_{jk}(v_i)}{g_{jk}} \tag{III.5}$$

where g_{jk} represents the total number of shortest paths between node v_j and node v_k, while $g_{jk}(v_i)$ is number of shortest paths that pass through v_i.

3.2.4.3 Closeness centrality

Closeness centrality is computed as the average shortest path between a node and any other node in the network. It measures how long it will take for information to flow from one node to the rest of the network. Closeness centrality can be computed as following;

$$Clos(v_j) = \frac{\sum_{j=1}^{N\setminus i} d_G(v_i, v_j)}{n-1} \tag{III.6}$$

where d_G represents the shortest path, or geodesic distance, between node v_i and node v_j

3.3 Network models

3.3.1 Random networks: Erdös-Rényi model

The first model for random networks was introduced by Erdös and Rényi [43]. For this model, each node in the network is connected to any other node with probability p. As a consequence, the average degree $\bar{k}$ is constant and the degree follows a Poisson distribution $P(k) \sim e^{-k}$ centered around the mean $\bar{k}$. Since a random network has a tree-like shape, the diameter of a random network is $D(G) \sim \frac{\ln n}{\ln \langle k \rangle}$. The clustering coefficient, instead, is computed as $C(G) = \frac{\#linksbetweennode}{\#maxnumberoflinks} = \frac{\frac{pk(k-1)}{2}}{\frac{k(k-1)}{2}} = p = \frac{\langle k \rangle}{n}$. Generally, random networks have low average path length and low clustering coefficient.

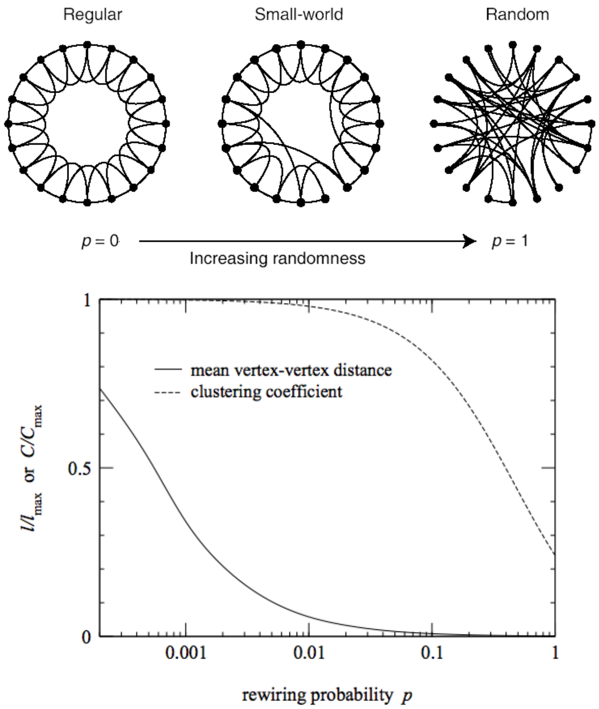
3.3.2 Small-world networks: Watts-Strogatz model

The so called “small-world” phenomenon was first discovered during the Milgram Experiment [44]. The objective of the experiment was to send a letter from a source person in Nebraska to a target person in Massachusetts. The first person was asked to send the letter to one of his/her acquaintances that was likely to be connected to the target person, with the objective of reaching the target within as few steps as possible. Milgram noticed that source and target were, on average, between five and six people apart, a number that was much lower than the number of people involved in the experiments, hence the term “six degree of separations”.

Later on, Watts and Strogatz [45] discovered that many real world networks, such as the brain network of the nematode species C. elegans, the Western U.S. power grid and the World wide web, even though of different kind, all have the same two properties: low average path length and high clustering coefficient. The network models known at that time, regular lattices and the random network model developed by Erdős and Rényi [43], fail to capture these properties. In fact, regular lattices have high average path length and high clustering coefficient, while random networks have low average path length and low clustering coefficient. They proposed a model that, starting from a regular lattice, randomly rewires edges according to a certain probability p between zero and one. If this probability is properly chosen, the model can generate small-world networks. In fact, they still preserve the high clustering coefficient of lattices, while the rewiring of few edges make the distance between nodes much lower. For these networks, the degree follows a Poisson distribution, the clustering coefficient is constant and the average path length has a logarithmic shape $L(G) = \log n$.

3.3.3 Scale free networks: Barabási-Albert model

Barabási and Albert [46] noticed that, for many complex networks, the degree distribution does not follow a Poisson distribution with a peak around the mean value, but rather a power-law distribution. This means that a very small number of nodes in the network (hubs) have a very high degree, something that the Watts-Strogatz model was still missing. They also realised that many real world networks show a preferential attachment: nodes do not connect randomly but favour more “popular” nodes. Furthermore, complex networks are not static and grow in size instead. They proposed a model that, based on these two



t

Figure 3.1: Watts-Strogatz model to generate small-world networks.

mechanisms, can generate networks with a power-law degree distribution. Networks are generated as follows:

- the network starts at $t = 0$ with n_0 nodes
- for every step, add a node with degree $k_i(i) = m_0$
- the new node is connected to other nodes in the network with probability proportional to their degree: $p(i) = \frac{k_i}{\sum_i^n k_i} = \frac{k_i}{2m_0t}$

The networks generated with this model are called scale-free networks. For these networks, the degree distribution follows a power law: $P(K) \sim k^{-3}$, the average path length is given by the formula $\langle APL \rangle \sim \frac{\log n}{\log \log n}$ and the clustering coefficient is given by $C(G) \sim n^{\cdot 0.75}$.

3.4 The emergence of communities

Another important property of complex networks is the organization into communities. Many real world networks such as social networks [47], the world wide web [35] and

Table 3.1: Degree distribution, average path length and clustering coefficient for different network models.

	Erdös-Rényi	Watts-Strogatz	Barabási-Albert
P(K)	$\frac{\lambda^k e^{-k}}{k!}$	Poisson	$\frac{2m_0^2}{k^3}$
APL	$\frac{\ln n}{\ln\langle k\rangle}$	$\log n$	$\frac{\log n}{\log\log n}$
C	$\frac{\langle k\rangle}{n}$	constant	$n^{.0.75}$

biological networks [36] show a community structure. A community consist of a group of nodes that are highly connected to each other and loosely connected to the rest of the network [47]. For example, researchers in a collaboration network tend to connect to other researchers in the same field, resulting in the emergence of communities that represent research topics.

3.4.1 Community detection

The objective of a community detection algorithm is to group similar nodes that belong to the same communities, i.e. that are more connected to each other than with the rest of the network. In the network partitioning problem, the network is divided into a fixed number of equally sized partitions. This problem is NP-hard [48]. Community detection is a generalization of the partitioning problem, where the number of communities to be found is not fixed and communities can vary in size. Therefore, the community detection problem is also NP-hard. Also, many community detection algorithms are based on modularity optimization. Modularity is a measure of partition quality of the network into disjoint communities [49] and finding the partition that maximizes modularity is known to be NP-Hard. Furthermore, modularity and other global network metrics used to find communities are not always accessible, since many real world networks such as the World Wide Web are too large to be completely known and very dynamic. Therefore, it is important that community detection algorithms maintain a low complexity and are highly scalable.

3.4.2 Performance metrics

In order to evaluate the performance of a community detection algorithm, as well as compare different algorithms, it is necessary to define some performance metrics.

3.4.2.1 Classical metrics

The most common metrics used to evaluate community detection algorithms come from classical clustering, where communities are seen as partitions of nodes. These partitions are compared to the ground-truth communities, when these are known.

Normalized Mutual Information (NMI) is an information theoretic metric that measures the amount of information that two partitions share. It ranges from 0 to 1, assigning 1 to communities that perfectly match the ground truth and 0 to a completely random assignment. The drawback of NMI is that it is affected by network size and number of communities in the network. For example, supposing that a community detection algorithm fails and assigns a different community to each node, the NMI will assume a value that is not the same for each network, hence impossible to interpret. The Adjusted *Randomized Index* (ARI) measures the proportion of pairs of nodes that are correctly assigned to the same community. It ranges from -1 to 1, assigning 1 for a perfect assignment, 0 for a random assignment and -1 for a bad assignment. Unlike NMI, it is not affected by any network characteristic.

Modularity measures the fraction of edges connecting vertices inside the same community and compares it to the same quantity computed on a random graph of the same size and average degree. Modularity will be higher if the network exhibits a community structure. It does not need the ground-truth community assignment to be known and is only based on the structure of the network. The drawback of modularity is the resolution limit: this metric is not accurate when computed on networks containing small communities. In order to solve this issue, many algorithms based on modularity optimization make use of a resolution limit parameter [50].

3.4.2.2 Topological metrics

There are also several metrics that allow to study the topological properties of a community assignment. The most common one is the *size* of a community. For many real world networks, the community size distribution follows a power-law, meaning that there is a majority of small communities and few large ones. The community size distribution, in general, provides very good information about the quality of a community assignment [51].

The *internal transitivity* of a community is defined as the average local transitivity over all nodes, where the local transitivity of a node measures the fraction of links between its neighbors. The formula is the following

$$\frac{1}{s_i}\frac{1}{k_i - 1}\sum_{h,j\in C}\frac{w_{i,j} + w_{i,h}}{2}a_{i,j}a_{i,h}a_{j,h} \tag{III.7}$$

where s_i is the strength of node i (sum all of the weights of its edges), k_i is the node internal degree, $w_{i,j}$ is the weight of the edge connecting nodes i and j and $a_{i,j}$ is an element of the adjacency matrix.

The *scaled density* is defined as the density of a community weighted by its size. The formula is the following:

$$\frac{2m_C}{n_C(n_C - 1)} \tag{III.8}$$

where m_C is the number of edges in the community and n_C is the number of nodes in the community.

The *average distance* of a community is the average shortest path between all pairs of nodes inside the community.

The *hub dominance* is defined as the maximal internal degree of a node divided by its maximum theoretical value. The formula is the following:

$$max_{i \in C}(\frac{k_i}{n_C - 1}) \tag{III.9}$$

Finally, the *internal modularity* is simply defined as the modularity of a community:

$$\frac{1}{2m_c} \sum_{i,j \in C} [(a_{i,j} - \frac{k_i k_j}{2m_c}) a_{i,j}] \tag{III.10}$$

These metrics have been shown to be a valid complementary tool to evaluate and compare community detection algorithms [52, 53]. Communities found by different algorithms can have similar NMI or ARI and still be topologically different. For example, the misclassification of a hub does not change the value of classical performance metrics, but may significantly affect the topological structure of a community For example, they may have different community size distributions.

3.5 Orbit analysis

Graphlets are small connected graphs of size between two and five. Graphlet analysis is a useful tool for analysing the global topological structure of networks and, locally, of a node's ego network. Figure 3.2 shows all the graphlets with up to four nodes. Some well known graphlets are the "star" graphlet and the "triangle" graphlet. Some graphlets are characteristic of certain type of network. For example, the triangle is more likely to be found in social networks, due to high transitivity, while the star graphlet is more likely to be found in visibility networks. Graphlet counts, defined as the number of times that each graphlet appears in a network, can be used to characterise networks.

Nodes within a specific graphlet can have different roles. For example, in the star graphlet, a node can be identified as the center and the other three nodes as the leaves.

Similarly, an orbit count can be defined as the number of times a node appear in each orbit, and can be used to identify group of nodes that cover different roles in the network. For example, the orbit count for the central position of the “brokerage” graphlet can be used to identify “mediator” nodes in collaboration networks.

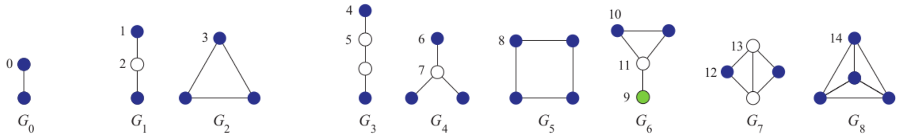


Figure 3.2: Graphlets with up to five nodes with their different orbits.

3.6 Exponential random graph models

Exponential random graph models (ERGMs) are a family of statistical models that help discover and understand the processes underlying network formation [20–22]. They have been used extensively in social network analysis and are popular in various fields such as sociology [54, 55], archaeology [56], and history [57]. ERGMs provide a model for a network that includes covariates, variables that relate to two or more nodes, which cannot be addressed using traditional methods. They can represent effects such as:

- *homophily*: the tendency of similar nodes, i.e. nodes having the same attribute, to form relationships.
- *mutuality*: the tendency of node B to form a relationship with node A, if node A is connected to node B.
- *triadic closure*: the tendency of node C to form a relationship with node A, if node A is connected to node B and node B is connected to node C.

ERGMs also provide maximum-likelihood estimates for the parameters governing these effects. For example, ERGMs can estimate the increased likelihood of a tie existing between two nodes when these nodes have the same attribute. ERGMs also provide a goodness-of-fit test for the model, in order to verify whether the effects included in the model are sufficient to explain the structure of the observed network. Furthermore, they can simulate networks that match the probability distributions estimated by the model. In other words, they can be used to generate artificial networks that reflect the characteristics of the observed network.

3.7 Temporal networks

In many real world applications, networks are not static. Nodes may appear and disappear at different times in the network. Similarly, edges may be active only for a limited amount of time [58]. For example, in a network of collaboration where actors are connected if they have worked on the same project, new actors are added to the network as they start their career and are removed from the network when they are no longer active. Similarly, edges between actors are active for the duration of the project only. In a co-authorship network, where actors are connected when they produce a certain output (i.e. scientific articles), the duration of the interaction can be considered negligible and a timestamp is associated to the connection (i.e publication date). In case of negligible duration of contact, a graph can be represented as $G = \{V, E, C\}$ where V represent the sets of vertices, E represents the set of edges and $C = \{t_1, \ldots, t_n\}$ represents a set of time contacts between vertices v_i and v_j. These graphs are called *contact graphs*. If edges represent continous interactions, instead, the graph can be represented as a set $G = \{V, E, T\}$ where $T = \{(t_1, t'_1), \ldots, (t_k, t'_k)\}$ represents the set of time intervals associated to E. These graphs are called *interval graphs*.

For both network types, an *aggregate graph* can be defined as a static graph where an edge between verteces v_i and v exist if there is any contact between them, regardless of the time. Furthermore, just like static graphs, temporal graphs can be undirected or directed, as well as unweighted or weighted.

Given a contact graph, it can be alternatively represented as a temporal sequence of static graphs, where the temporal graph is “sliced” into several static graphs that include all edges with the corresponding timestamp. If the time resolution is too high, these graphs may result too sparse, therefore a time window can be used instead, where each time slice includes all edges with a timestamp that falls within such time window. This time windows can be disjoint or overlapping. Disjoint time windows can be used such that edges are represented in one time slice only, but this may result in graphs being too different between each other at consecutive time slices. Overlapping time windows, instead will allow them to appear in multiple contiguous time slices, in order to have a more gradual change over time. The same concept of slicing and time windows can be applied to interval graphs. In this case, each time slice will include all edges whose time interval overlaps, either completely or partially, with the time windows. Again, a complete overlap may cause graphs in consecutive time slices to be very sparse and too different from each other, while a partial overlap would allow more continuity.

Finally, time windows size can also be adapted such that each time slice contains and homogeneous number of nodes and/or edges.

CHAPTER IV

MEMLPA: A MEMORY-BASED LABEL PROPAGATION ALGORITHM

This chapter is partially based on the published articles “A memory-based label propagation algorithm for community detection” [59] and “Local memory boosts label propagation for community detection” [60]. In this chapter we introduce MemLPA, a community detection algorithm that is based on the label propagation algorithm and incorporates a memory element, in order for nodes to consider past states of the network in their decision rule.

4.1 Community detection algorithms

Girvan and Newman [61] were first to propose a divisive hierarchical algorithm based on edge betweenness: given an edge, it measures the number of shortest paths between all pairs of nodes in the network that pass through this edge. Removing edges with high betweenness will enhance the separation of communities. This method ranks edges according to their betweenness and iteratively removes them. At the end, the configuration that achieves the highest modularity is chosen. Its complexity is $O(nm^2)$. A faster version of this method was also proposed [62]: it is a heuristic algorithm that, at each iteration, merges nodes into communities to optimize modularity. This method runs in $O(md \log n)$, where *d* is the depth of the dendrogram. Blondel, Guillaume, and Lambiotte [63] proposed a similar method called Louvain. All nodes are initially assigned to a different community and, at each iteration, each node is moved to the community that achieves the highest modularity improvement. Once communities are defined, a new network is built, where nodes represent the communities found. The process iterates until improvement no longer occurs. It runs in $O(n \log n)$.

Walktrap [64] defines a similarity between nodes according to the transition probability of random walkers. A random walker is an agent that, starting from a random node, moves from one node to another with a uniform probability. It runs in $O(n^2 m)$ or $O(n^2 \log n)$ on sparse networks. Infomap, similarly, is a global optimization method that optimizes a quality function defining the code length of a random walk process in the network. Its complexity is $O(m)$. Newman [65] also proposed a spectral method based on the Eigenspectrum of the modularity matrix. Its leading eigenvector is computed and the network is split into two sub-communities such that modularity is maximized. The process is then repeated on the communities just found. This method runs in $O(n(m+n))$ or $O(n^2)$ on sparse networks. Finally, Reichardt and Bornhol [66] interpreted community detection as the minimization of the energy function of a spin model, where communities are seen as spin configurations. It runs in $O(n^{3.2})$ on sparse networks.

4.2 Label propagation algorithm

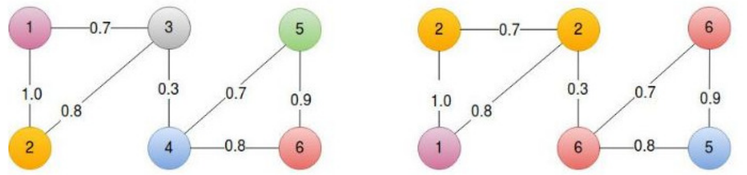
Many of the algorithms described are not suitable for large-scale networks: they have high complexity and require global information of the network. To overcome this problem, Raghavan [67] proposed the Label Propagation Algorithm. It initially assigns a distinct label to each node. Labels are then iteratively updated following the majority voting rule, until consensus is reached among all nodes in the network. This method runs in linear time, is scalable and uses the network's local information only, without the need of optimizing any objective function. Unfortunately, LPA gets easily stuck in local optima and is thus outperformed by more recent and sophisticated algorithms. Furthermore, in some cases a certain label may over-propagate and create a single giant community.

In order to overcome these issues, several improvements have been proposed. Clark [68] developed a variation of LPA that takes into account modularity when applying the majority rule. This method was extended by Liu and Murata [69] with a greedy method that, given the communities found, merges them in an attempt to improve modularity, allowing the algorithm to escape from local optima. Leung [70] introduced a decision rule based on node preference, in this case node degree, to improve performance: when a node applies the decision rule, labels of nodes having a higher degree will be assigned a higher score. They also extended the algorithm with hop attenuation: every time a label is propagated through the network, a negative score is assigned to it in order to prevent a certain label from flooding the network. The algorithm is scalable and still runs in linear time. Xie and Szymanski [71] proposed another node preference, based on neighborhood overlapping, that is shown to be related to the clustering coefficient. Šubelj and Bajec [72] elaborated two particular strategies, called defensive preservation and offensive expansion, that adapt node preference to focus on core nodes and border nodes of communities. They are combined and applied hierarchically. They also found that the network structure affects the effectiveness of node preference and hop attenuation. This algorithm runs in $O(m^{1.19})$ and is highly scalable. Xie and Szymanski [73] also developed LabelRank, a variation of the classical LPA that takes inspiration from the Markov Cluster Algorithm (MCL) [74]. Instead of a single label, each node maintains a list of label distributions that is updated at each iteration. An inflation operator is used to enhance the gap between strong and weak labels, while a cutoff operator is applied to remove labels below a certain threshold, in order to shorten these lists and make the computation more efficient.

To our knowledge, there are only few methods that explicitly refer to the use of memory in LPA, where nodes collect labels from previous iterations to keep track of past states of the network. The Speaker-Lister Label Propagation Algorithm (SLPA) [75] is based on an information dynamic rule: for each node, its neighbors select one label from their memory according to a speaking rule and the node updates its memory according to a listener rule. After a fixed number of iterations, a thresholding procedure is applied to each node's memory to assign it to one or multiple communities. Another memory-based LPA algorithm (MLPA) [76] implements a memory element in each node where, at each iteration, the label chosen is stored. In this way, it is possible to know the community structure of the network at each iteration and have snapshot of the evolution of the network. After a fixed number of iterations, the most frequent label in each node's memory is chosen and a last round of the classical LPA is performed to assign nodes to communities. Finally, a more recent LPA variation called Fluid Communities (fluidC) [77], is based on the idea of fluids expanding and contracting as a result of their interaction. The algorithm initializes a certain number of community seeds in the network and each node updates its label using an update rule based on fluid density. This algorithm requires the number of communities to be set at start. Some work on consensus dynamics also refers to memory: a non-deterministic version of the Naming Game [78,79], which is similar in some aspects to LPA, extends the agents with local memory.

4.3 MemLPA: a memory-based label propagation algorithm

In the classical LPA, each node updates its label according to the current state of the network. Each node collects its neighbors' labels and selects the most chosen one according to a majority rule. This mechanism does not consider past states of the network, since each node collects new labels at each iteration and discards the previous ones, making the algorithm memory-less. In this section we introduce MemLPA, a variation of the classical



Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
2 – 1.0 3 – 0.7	**1 – 1.0** 3 – 0.8	1 – 0.7 **2 – 0.8** 4 – 0.3	3 – 0.3 5 – 0.7 **6 – 0.8**	4 – 0.7 **6 – 0.9**	4 – 0.8 **5 – 0.9**

Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
2 – 1.7 3 – 0.7 1 – 1.0	1 – 1.0 3 – 0.8 **2 – 1.8**	1 – 1.5 **2 – 1.5** 4 – 0.3 6 – 0.3	3 – 0.3 **5 – 1.5** 6 – 1.5 2 – 0.3	4 – 0.7 **6 – 1.6** 5 – 0.9	4 – 0.8 5 – 0.9 6 – 1.7

Figure 4.1: Iterations of MemLPA on a weighted undirected graph. Color and node number represent labels. Columns in the table represents nodes memory. In the first iteration, node 1 receives labels 2 and 3 from its neighbors. Using the classical LPA, in the second iteration, it would discard these labels and collect new ones (label 1 and 2). With memory, instead, the old labels are not discarded but updated, therefore node 1 contains labels 1, 2 and 3.

LPA where nodes implement a memory mechanism that allows them to "remember" about past states of the network and uses a decision rule that takes this information into account.

4.3.1 Algorithm description

When using memory, labels are not discarded but updated at each iteration. Each node maintains a list of labels with its associated score. Initially, each node is assigned a distinct label (line 2 of the pseudo-code) and its memory is empty (line 3). At each iteration, each node collects its neighbors' labels (line 6) and updates its memory according to edge weight (for weighted networks) and node preference (line 7). If a new label is not in memory already, a new entry is created, otherwise the score for the corresponding label is updated. Each node then selects a label from its memory using a decision rule that takes into account the labels' score, in this case the label having maximum score (line 9). This mechanism can be applied to directed or undirected as well as to weighted or unweighted graphs. Figure 4.1 shows how MemLPA works.

In order to keep MemLPA scalable, a synchronous update rule is used: each node independently updates its label according to the state of the network during the previous iteration. In fact, a synchronous update may cause LPA to oscillate between two different configurations. Section 4.4 shows how the two different update rules affect the convergence of MemLPA. As node preference, we use a heuristic based on neighborhood overlapping, computing the fraction of neighbors that a node shares with another. When updating a node's memory, a higher score will be assigned to labels coming from nodes that have many neighbors in common. Section 4.4 shows the impact of this heuristic on performance. To speed up the algorithm, we define a cutoff operator to prune each node's memory (line 8). At each iteration, all labels below a certain threshold are deleted, keeping only the most relevant ones. Regarding the termination criterion, several options have been proposed in the literature, based on convergence, modularity improvement, active nodes and scarcity of updates. Many of these options are based on global information of the network, therefore we decided to use a termination criterion based on active node list: a node is considered active if the label chosen during the current iteration is different from the previous one or if any of its neighbors becomes active again. The active node list initially contains all nodes (line 4) and at each iteration a node is removed if it is no longer active or it is added if it becomes active again (line 10). The decision rule is applied only on active nodes and the algorithm terminates when the active node list is empty. This keeps the algorithm decentralized and speeds up the algorithm compared to applying the decision rule on every node. Section 4.4 shows how the termination criterion based on active node list affects performance and convergence of the algorithm. The decision rule based on memory that MemLPA uses, as well as MLPA, may result in singleton communities. For these nodes, an additional round of label propagation without memory is performed in order to assign them to a bigger community.

Algorithm 1 MemLPA

```
Input  : Graph G(N, E)
Output: Communities C
for $n \in N$ do
    $c_n \leftarrow l_n$ //Assign unique label to nodes
    $M_n \leftarrow \emptyset$ //Initialize memory
$AL \leftarrow N$ //Initialize active list
while $AL \neq \emptyset$ do
    for $n \in AL$ do
        $C_n \leftarrow CollectLabels(Neigh(n))$
        $M_n \leftarrow UpdateMemory(C_n)$
        $M_n \leftarrow \{l_n^m \in M_n, m \in N \mid |mean(M_n) - sd(M_n)| \leq l_n^m\}$
        $c_n \leftarrow ApplyRule(M_n)$
    $AL \leftarrow UpdateActiveList(AL)$
```

4.3.2 Computational complexity

The computational complexity of MemLPA on a certain node, where k is the average degree and h is the average memory length, can be assessed this way:

- Collecting labels for a node with k neighbors has complexity $O(k)$.
- Updating a node's memory with k new values has complexity $O(k)$.
- Using the cutoff operator on a node's memory has complexity $O(k)$.
- Choosing a new label from memory has complexity $O(h)$.

Node preference, if used, can also affect complexity. Neighborhood overlapping, on a node with k neighbors, has complexity $O(k)$ [71], while node preference based on node degree has complexity $O(1)$ [70]. Notice that the information needed for node preference must only be computed during the first iteration and nodes can store and reuse this information. In section 4.4 we show how the cutoff operator keeps the average memory length constant and significantly lower than the average node degree. Iterating on all nodes, the overall complexity of MemLPA is $O(k*n)$ or $O(m)$, therefore comparable to $O(m)$ of the classical LPA [67]. Therefore, the complexity of MemLPA is still linear with respect to the number of edges in the network.

4.4 Results

We implemented MemLPA and assessed the use of memory and some of the variations proposed in the literature. We then compared it to other memory-based label propagation algorithms and well-known community detection algorithms. We also ran MemLPA to study some of its characteristics that are important for the convergence of the algorithm. For the analysis we ran all algorithms on the LFR benchmark [80], an established benchmark in the literature for community detection, that allows to generate networks with properties similar to real world networks. As performance metrics, we used classical clustering metrics such as Normalized Mutual Information (NMI) [17, 18] and Adjusted Rand Index (ARI) [19], as well as topological metrics such as community size, internal transitivity, scaled density, average distance, hub dominance and internal modularity [52]. We also applied these algorithms on a set of real world networks of different nature and used the modularity measure [61] to evaluate the quality of the community assignments found.

4.4.1 Cluster analysis

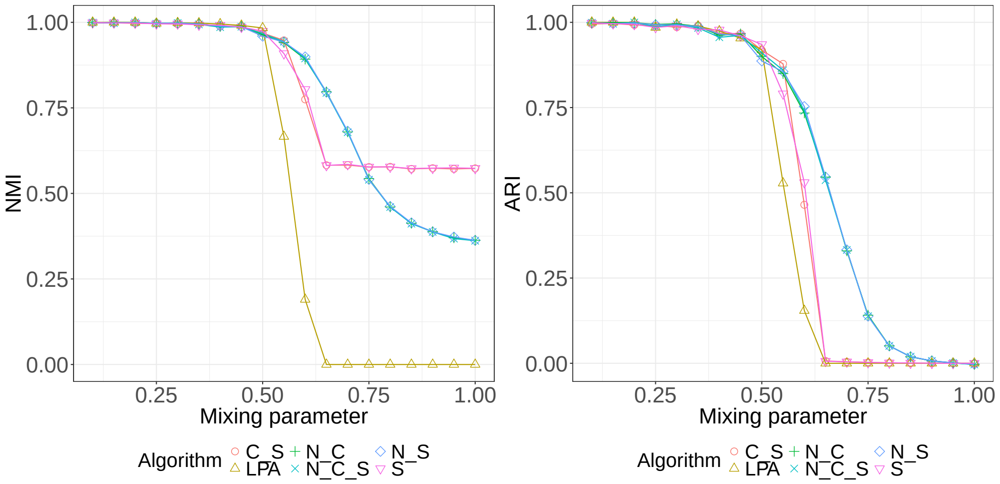
In this section we use classical clustering metrics such as NMI and ARI to assess the use of memory and some of the variations proposed in literature. We then compare MemLPA to other memory-based label propagation algorithms. Finally, we compare MemLPA to other well-known community detection algorithms. We also run MemLPA to study its convergence. We run all algorithms on the LFR benchmark and a set of real world networks.

4.4.1.1 Artificial networks

The first set of experiments was conducted on the LFR benchmark to investigate the advantages of the LPA variations chosen and the use of memory. A mixing parameter μ controls the portion of intra-community edges. Node degree and community size distribution, like in many real world networks, follow a power-law distribution. Benchmark graphs were generated with a number of nodes *N* = 1000, minimum community size *C.min* = 10, maximum community size *C.max* = 50, average degree *K.avg* = 20, maximum degree *K.max* = 50, degree exponent *K.exp* = 2 and community size exponent *C.exp* = 1, while μ was dynamically changed.

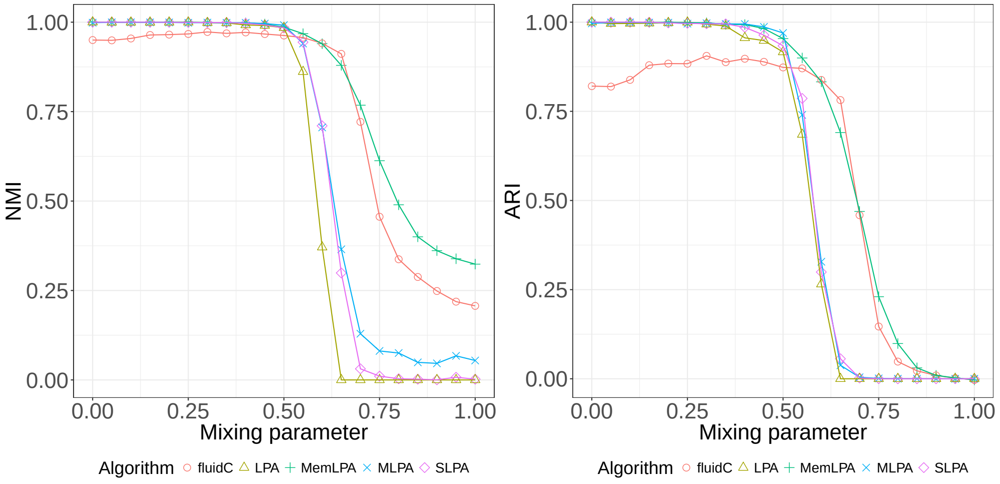
We compared the classical LPA to different variations of MemLPA that use synchronous (S) and asynchronous update rule, with and without node preference (N), with and without cutoff operator (C). Figure 4.2 shows that using a synchronous or asynchronous update does not make a significant change in performance (N_C_S vs N_C). Using the cutoff operator does not degrade performance either (C_S vs S and N_C_S vs N_S). This shows that MemLPA can be decentralized and scalable without any loss in performance. For low values of μ all variations obtained optimal results. The classical version of LPA, the only one not using memory, was the first algorithm to drop in performance for $\mu \geq 0.5$. In fact, a label flooded the network and created a single giant community. This confirms that the use of memory improves performance and prevents a label from over-propagating in the network. For $\mu \in [0.5, 0.7]$ the variations that use node preference (N_S, N_C and N_C_S) obtained the best results, but it is not the case for higher values. In fact, the variations that did not use node preference (S and C_S) obtained higher values of NMI for $\mu \in [0.7, 1]$. We must consider that the NMI depends on network size and number of communities. Therefore we decided to look at the ARI to have a more accurate comparison. We can see that, in this case, variations using node preference actually achieve better results.

The same experiment was conducted to compare MemLPA to other memory-based label propagation algorithms. Figure 4.3 shows that, for low values of μ, LPA, MemLPA and SLPA achieve perfect results, while it's not the case for fluidC and MLPA. For $\mu \geq 0.5$, all algorithms' performance start dropping. LPA's performance is first to drop to zero,



(a) (b)

Figure 4.2: Experiments on the LFR benchmark. All variations, except the classical LPA, implement memory. N: node preference, C: cutoff operator, S: synchronous update. Experiments are run 20 times and results averaged.

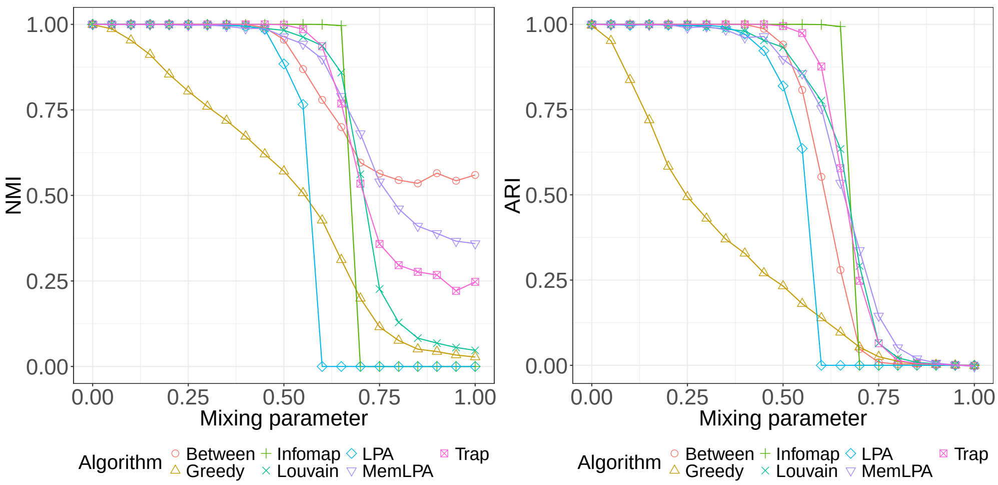


(a) (b)

Figure 4.3: Experiments on the LFR benchmark. Experiments are run 20 times and results averaged.

showing that the use of memory in any of the algorithms is beneficial. LPA and SLPA both find a single giant community, while MemLPA achieves the best performance overall.

Finally, we compared MemLPA to other well-known community detection algorithms. We chose some of the algorithms described in section 4.1 (all available in the



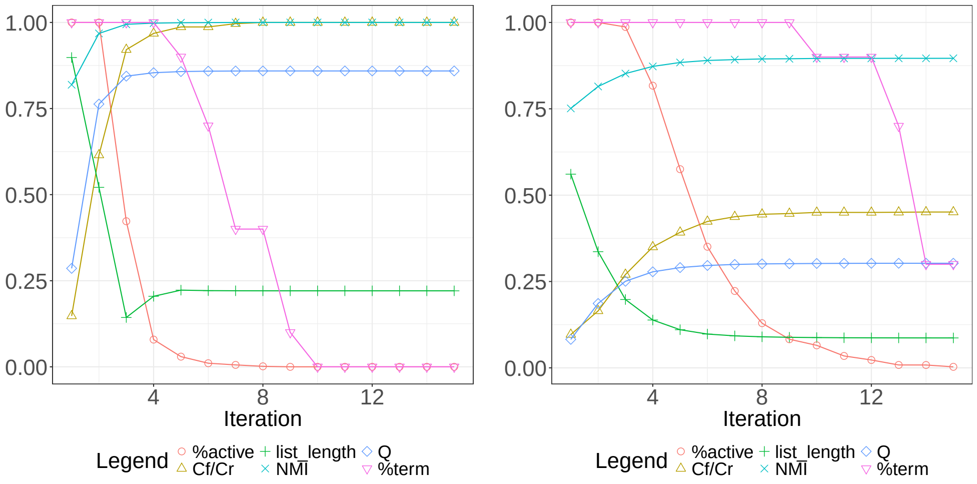
(a)

(b)

Figure 4.4: Experiments on the LFR benchmark. Experiments are run 20 times and results averaged.

igraph R package [81]). Figure 4.4 shows that, for low values of μ, most algorithms obtain optimal results, while Greedy gradually decreas in performance. For $\mu \in [0.5, 0.7]$ most of the algorithms start degrading in performance, especially LPA and Between. MemLPA, in this range, is only outperformed by Infomap and Trap. For $\mu \geq 0.7$ MemLPA is the best algorithm after Between but, looking at the ARI, MemLPA performs slightly better until all algorithms' performance drop.

We also conducted two experiments to analyze some of the characteristics of MemLPA at run-time. We used $\mu = 0.1$ to generate networks where communities are very well defined, and $\mu = 0.6$ for loose communities. As performance measures we recorded NMI, modularity and the ratio between the number of communities found by MemLPA and real communities. The information that we recorded is the percentage of runs that terminated, the number of active nodes and the average ratio between memory length and node degree. Figure 4.5 shows that, for $\mu = 0.1$, MemLPA increases in performance quickly, being able to find the correct number of communities. The percentage of active nodes drops significantly right after the best performance was reached, causing most of the runs to terminate.



(a) (b)

Figure 4.5: Experiments on the LFR benchmark. $\mu = 0.1$ (a) and $\mu = 0.6$ (b) has been used for the two experiments. On the x-axis you can find number of iterations. On the y-axis NMI, modularity, ratio between number of communities found by MemLPA and real communities, percentage of runs that terminated, number of active nodes and average ratio between memory length and node degree. Both experiments have been run 50 times and results averaged.

The average memory length drops significantly during the first iterations and then stabilizes, holding a constant value that is significantly lower than average node degree. For $\mu = 0.6$, as expected, there is a similar behavior but the algorithm is slower to converge. Surprisingly, the average memory length is lower for $\mu = 0.6$. A possible explanation is that nodes in well defined communities hold very strong labels in their memory, while for loose communities labels are weaker and more likely to be removed by the cutoff operator.

4.4.1.2 Real world networks

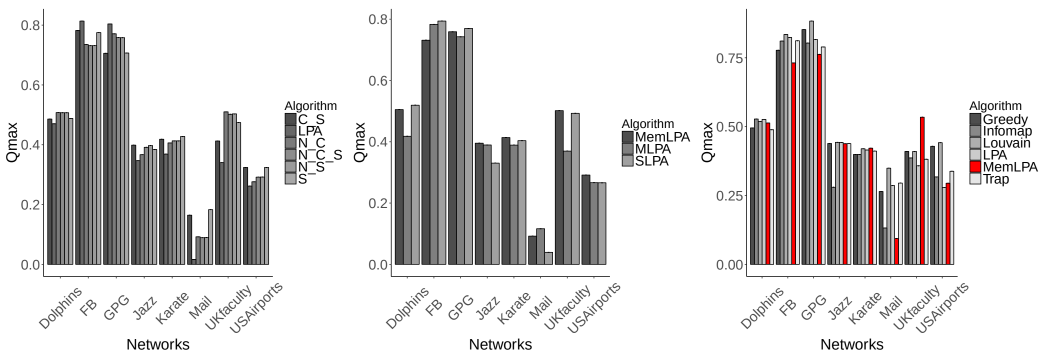
We conducted similar experiments on a set of real world networks of different nature. An overview of these networks characteristics is provided in Table 4.1.

In the first experiment, similarly to Section 4.4.1.1 for artificial networks, we investigated the advantages of memory and the LPA variations chosen. Figure 4.6 shows that using a synchronous or asynchronous update did not make a significant change (N_C_S vs

Table 4.1: Real world networks characteristics.

	#nodes	#edges	directed	weighted
karate	34	78	no	yes
UKfaculty	81	817	yes	yes
mail	184	2116	yes	no
dolphins	62	159	yes	no
jazz	198	2742	yes	no
USAirports	755	23473	yes	yes
FB	4039	88234	no	no
PGP	10680	24340	no	no

N_C) and the cutoff operator did not degrade performance (C_S vs S and N_C_S vs N_S). This allows MemLPA to be scalable, fast and performing. Node preference did not affect performance on unweighted networks significantly, while performance mostly degraded for the weighted ones (N_S vs S and N_C_S vs C_S). A possible explanation is that weight is a more significant factor than neighborhood overlapping when it comes to measuring the similarity between nodes. Additionally, other types of heuristics might be more effective, such as node degree. Implementing memory was beneficial on most networks when compared to the memory-less LPA. In particular, it prevented labels from over-propagating on the *Mail* network where the classical LPA finds a giant community that contains about 95% of the nodes and few very small ones. The only case where the classical LPA obtained better results is for unweighted and undirected networks (*FB* and *PGP*). In the second experiment we compared MemLPA to other memory-based label propagation algorithms. fluidC was not considered since it requires the number of communities as input. MemLPA achieved the best performance on *Jazz*, *Karate*, *UKfaculty* and *USAiports* networks, while still obtaining good results on *Dolphins* and *GPG* networks. Finally, we compared MemLPA to other well-known community detection algorithms. MemLPA was among the most performing algorithms on all networks, obtaining the best results on *Karate* and *UKFaculty* network. Again, MemLPA did not obtain optimal results for unweighted and undirected networks. It must be underlined that modularity may not be an optimal metric, because



(a) (b) (c)

Figure 4.6: Experiments on real world networks. Each bar plot represents the results obtained for all algorithms on a specific network. Experiments have been run 100 times and results averaged.

of the resolution limit and the fact that networks may present different community scales. Also, using different resolution limit parameters can affect the results.

4.4.2 Topological analysis

The topological analysis was conducted on the LFR benchmark as supplementary evaluation. We decided to focus on specific values of μ for which the classical LPA starts to fail to identify communities. Benchmark graphs were generated with a number of nodes $N = 1000$, minimum community size $C.min = 10$, maximum community size $C.max = 50$, average degree $K.avg = 20$, maximum degree $K.max = 50$, degree exponent $K.exp = 2$, community size exponent $C.exp = 1$ and mixing parameter $\mu \in [0.55, 0.6]$. Each algorithm was run on each instance of the benchmark. For each community found, all topological metrics presented in 3.4.2.2 were computed and the results averaged on communities having the same size. In order to quantify the agreement between the ground truth and the communities found, we performed a Kolmogorov-Smirnov (KS) test, used to test if two samples are drawn from the same distribution. The KS distance between the two distributions is then computed for each algorithm.

We compared the classical LPA to variations of MemLPA with and without node preference (N), using the ground-truth community assignment as reference. Figure 4.7, and the KS distance computed between the ground truth and the communities found, show that using memory is beneficial for community size, internal modularity, internal transitivity and scaled density. The classical LPA performs better only for average distance and hub dominance.

The same experiment was performed to compare MemLPA using node preference to other memory-based label propagation algorithms. MLPA was not considered since it generates many disconnected communities and singleton communities for which most of the metrics cannot be computed. Results are shown in Figure 4.8. MemLPA achieves best results for community size and hub dominance, and second best results for internal transitivity and average distance.

Finally, we compared MemLPA using node preference to some of the well-known community detection algorithms presented in section 4.4.1.1, they achieved similar results. Figure 4.9, and the KS distance, show that MemLPA achieves the second best results only for the average distance.

MemLPA finds a greater number of smaller communities, which affects the quality of the communities found. When choosing a label from memory, a label that was very frequent in the first iterations but not as much in the last ones will be still selected. A node may select a label from one of the first iterations and a neighboring node a label from the last iterations. As a consequence two smaller communities will form instead of a single bigger one. Finally, for the topological properties, the use of node preference based on node overlapping is not always beneficial, compared to the classical decision rule.

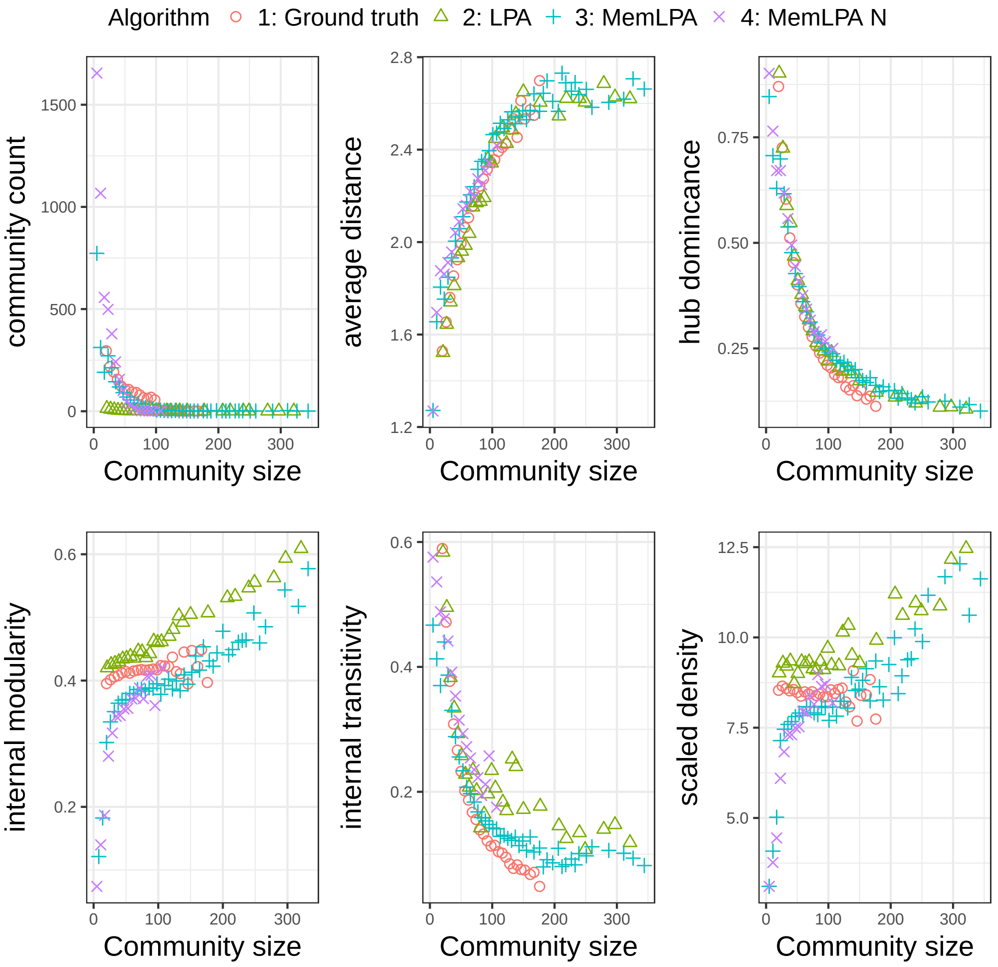


Figure 4.7: Experiments on the LFR benchmark. N: node preference. Experiments are run 500 times and results averaged on communities having same size. The KS distance between the ground truth and communities found is shown in the table below.

	LPA	MemLPA	MemLPA N
community size	0.54	0.42	0.34
average distance	0.21	0.60	0.48
hub dominance	0.14	0.50	0.55
internal modularity	0.81	0.60	0.93
internal transitivity	0.47	0.32	0.69
scaled density	0.93	0.63	0.78

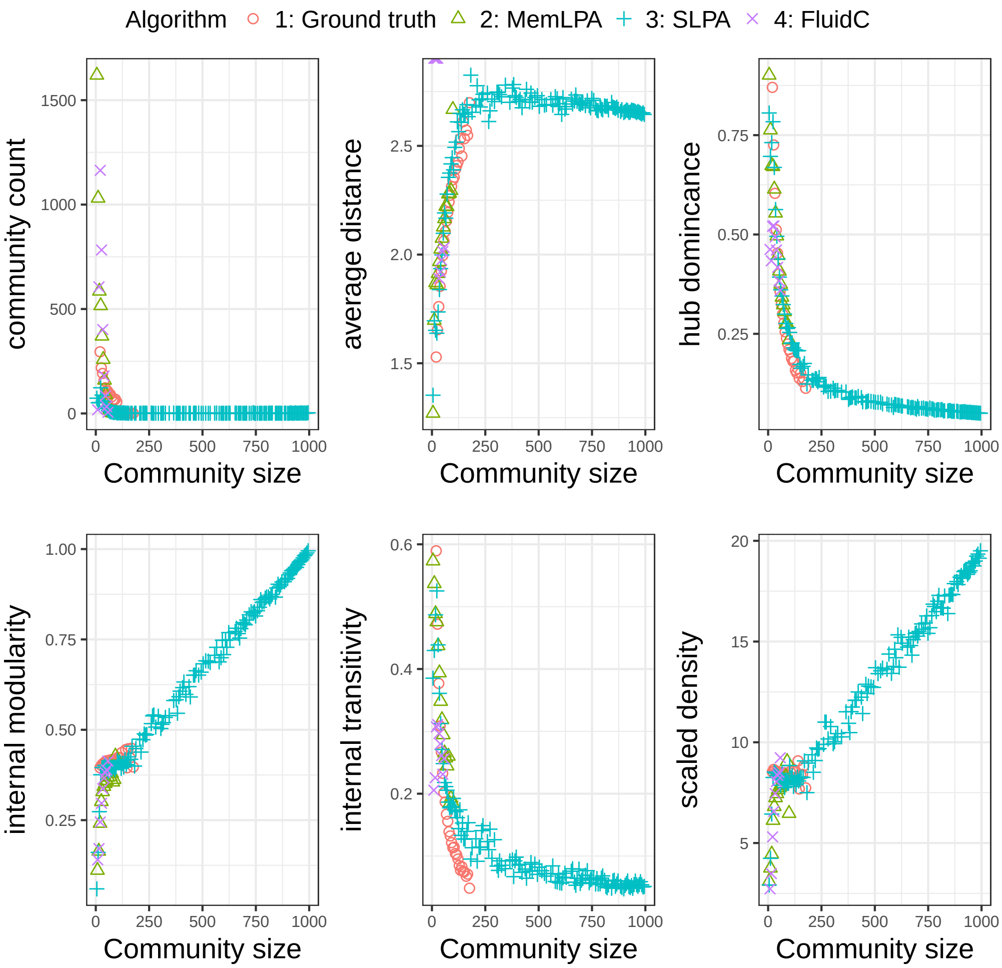


Figure 4.8: Experiments on the LFR benchmark. N: node preference. Experiments are run 500 times and results averaged on communities having same size. The KS distance between the ground truth and communities found is shown in the table below.

	MemLPA N	SLPA	FluidC
community size	0.34	0.51	0.46
average distance	0.48	0.79	0.44
hub dominance	0.55	0.69	0.77
internal modularity	0.93	0.76	0.89
internal transitivity	0.69	0.38	0.77
scaled density	0.78	0.75	0.59

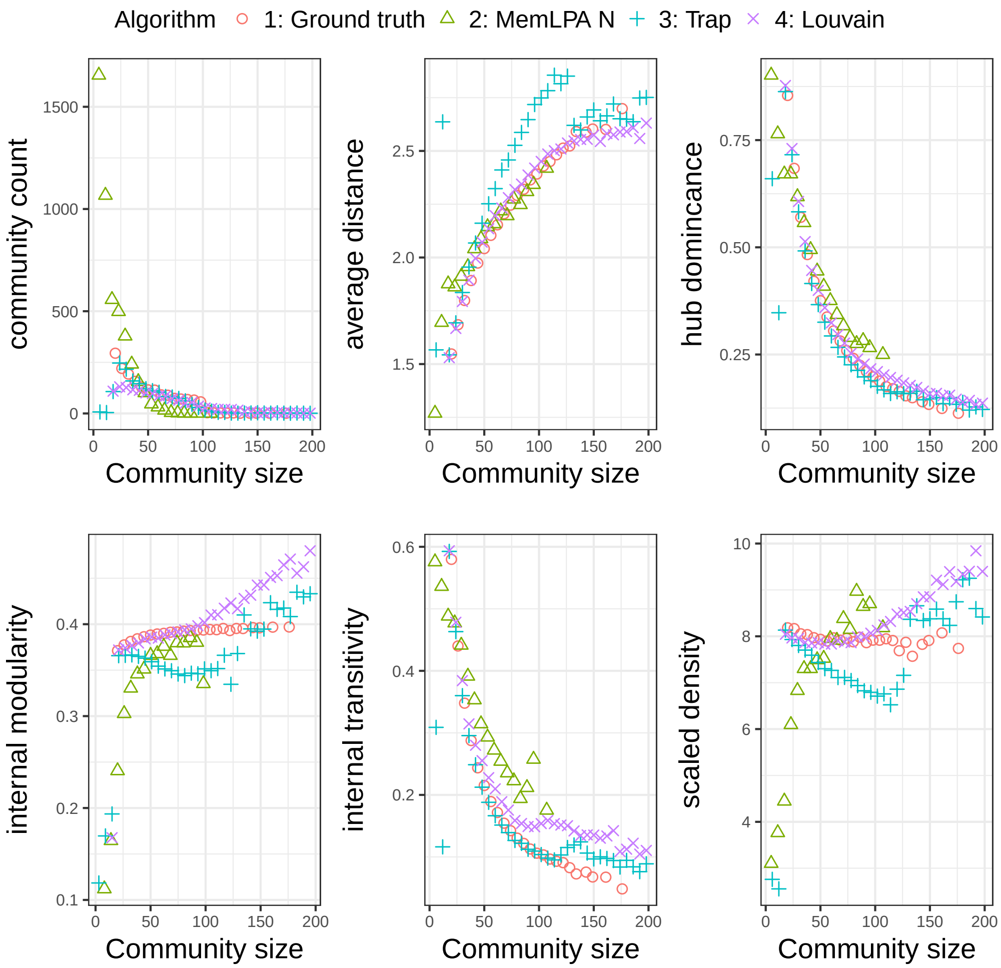


Figure 4.9: Experiments on the LFR benchmark. N: node preference. Experiments are run 500 times and results averaged on communities having same size. The KS distance between the ground truth and communities found is shown in the table below.

	MemLPA N	Trap	Louvain
community size	0.34	0.23	0.23
average distance	0.48	0.53	0.22
hub dominance	0.55	0.19	0.12
internal modularity	0.93	0.67	0.58
internal transitivity	0.69	0.22	0.43
scaled density	0.78	0.48	0.52

4.5 Conclusion

In this study we proposed MemLPA, a variation of LPA where nodes implement a memory mechanism that allows them to “remember” past states of the network and use a decision rule that takes this information into account. It runs in linear time, is scalable and only uses local information of the network. We gave an overview on community detection algorithms, LPA and the LPA variations proposed in the literature. We investigated the advantages of memory and we found that its usage increases performance and prevents labels from over-propagating over the entire network, resulting in a single huge community. We conducted extensive experiments on the LFR benchmark and used NMI and ARI as performance metrics. We tested MemLPA against other existing label propagation algorithms that implement memory to show that it provides better results. We also compared MemLPA to well-known community detection algorithms to show that it outperforms some of them for values of the mixing parameter between 0.5 and 0.8. Then, we conducted experiments on a set of real world networks of different nature, using modularity to evaluate the quality of the community assignments found, that further confirmed our findings. Finally, we performed a topological analysis using the LFR benchmark, comparing the topological properties of the communities found to the ground-truth community structure.

CHAPTER V

MINORITIES IN COMPUTER SCIENCE. GENDER AND ETHNIC COLLABORATION PATTERNS IN A TEMPORAL CO-AUTHORSHIP NETWORK

This chapter is partially based on the article “Minorities in computer science. Gender and ethnic collaboration patterns in a temporal co-authorship network” [34], currently under review.

In previous research [82] it was demonstrated the gender imbalance at computer science (CS) conferences by research area, which was identified through topic modeling and grouped using a clustering algorithm. The data [83] showed consistently low female authorship at 18,4%, with less interdisciplinary and more specialist research areas exacerbating gender inequality in favour of men [82]. The objective for this work is to further analyse the social network dynamics of computer science researchers and their collaboration patterns, expanding the analysis to include ethnic diversity, or rather a lack thereof.

Contrary to the current gender balance in computer science, women occupied prominent positions during the pioneering years of the history of computing in 1950s and many started entering the computing profession from the mid-1960s onward. As Tom Misa explains in the introduction of *Gender Codes* [84], “women earned 37% of all U.S. bachelor degrees in computing” at the peak of their involvement in computer science in 1984. Since then, this trend has been reversed and women have been leaving the field. One of the barriers for female graduate students was the added requirement of previous programming experience for computer science programs, as well as any other general action taken to thin the number of students enrolled. Furthermore, certain gender specific topics used for assignments such as sports data, as well as mentoring programs dedicated to male students only, have not encouraged female students to pursue and obtain a degree in computer science. Marketing has also played a critical role. Advertisements of computer games in the early 1980s reinforced the “computer geek” stereotype that sees only men interested in computers and video games.

Although computer science and STEM in general is often thought to be merit-based, academic institutions are not neutral spaces where race, ethnicity and gender have no effect [85]. In this work, gender is considered as ’culturally based as opposed to biological sex differences’ [86]. In the same fashion, ethnicity denotes ’groups that share a common identity-based ancestry, language, or culture’ [87]. In computer science, women and faculty of colour are in the minority and their experience in academia differs from their white male colleagues [88, 89]. Academia is not a meritocracy [90], instead it forces the minority ’to conform to the status quo of the dominant group’ [85].

Several approaches to quantifying this gender and ethnic gap has been used, such as bibliometric analysis, statistical modeling and social network analysis. Agarwal et al., in their bibliometric analysis, present empirical evidence of the under-representation of minorities in computer science for both gender [91] and ethnicity [92]. In their study on ethnicity [92], they found that most articles in computer science are published by East-Asian, British and Indian researchers, whereas French, African and Nordic researchers published least. Furthermore, East-Asian researchers have the highest representation of women at between 29,89% and 36,77% [92]. A very extensive bibliometric analysis [93] has recently confirmed that women are under-represented in most scientific disciplines, publish fewer articles during their career, have shorter career length and higher dropout rate. AlShebli, Rahwan, and Woon show the benefits of diversity, in terms of gender and ethnicity, to scientific collaboration [94]. Finally, Aspray provides an overview on STEM education for women and ethnic minorities, followed by several case studies of organizations that successfully promoted women participation in Computer Science [95].

Social networks are defined as a collection of individuals that are connected to each other if they form some kind of relationship, like friendship, acquaintanceship or collaboration. Social Network Analysis

> is concerned with the patterns formed by the [nodes] and [edges] and involves exploring these patterns, mathematically or visually, in order to assess their effects on the individuals (...) that are the members of the 'networks'. [4]

A co-authorship network is a special class of social network, where nodes represent authors that are connected if they have co-authored one or more papers [96]. Starting around year 2000, thanks to the availability of online bibliographies, researchers were able to build large co-authorship networks and interest in this topic has grown ever since [97–103]. Using social network analysis techniques, M. Jadidi et al [104] show that women are less likely to adopt the collaboration patterns that lead to success, have sparser ego networks and show higher gender homophily. Bravo-Hermsdorff et al. [105] find a steady increase in participation by women and define some network metrics to measure the structural importance of an authorship, showing a substantial difference between the collaboration patterns of women and men.

What is still missing is equally extensive research on the under-representation of ethnic minorities in computer science research. This work aims to fill that void by analysing collaboration patterns not only based on gender, but also based on ethnicity.

5.1 Methods

Most metrics commonly used to quantify the success of a researcher are based on counts of number of papers published in their career and number of citations [97–100,104]. At the same time there are other alternatives, based on social network analysis techniques, that look at the position of researchers in the co-authorship network and patterns of collaboration [97–99,101,103–105]. Social network analysis can be used to indicate a researcher's chances of success based on their position in the academic network, rather than evaluating productivity in terms of number of publications or citation score. This work mainly tries to unveil how gender and ethnic disparity has shaped the co-authorship network in computer science over the past 55 years.

5.1.1 Data set

The data set is publicly available at https://data.mendeley.com/datasets/3p9w84t5mr/1 and contains 112 456 papers written by 126 094 authors that were published at 81 different CS conferences between 1960 and 2015 [83]. The data set resulted from a snapshot of the DBLP bibliographic database taken on 17 September 2015 [83]. Papers indicate who is connected in the co-authorship network. However, papers are not identified by title, but by their unique URL. As a result, all papers with URL listed as 'unknown' had to be filtered. A co-authorship network was created using this data set. Since many network metrics are only defined for connected networks, only the giant component was considered, which is the biggest connected component in a network.

5.1.2 Gender

For this work, gender was considered as 'culturally based as opposed to biological sex differences' [86]. The data set includes author gender, which was generated by the Genderize API [106] based on the first name of an author. At the time the data set was created, Genderize API based results on a data base containing “2 162 860 distinct names across 79 countries and 89 languages” [91]. Based on first name, the algorithm then determines whether an author is male, female or not assigned, including a confidence score. All authors with a gender accuracy below 99% were filtered out.

As a result of this filter, the data set used contains 78 conferences, 98 912 papers and 69 087 authors. Therefore, 34% of the authors were left out and, considering that there are 3,11 authors per paper on average, an estimate of 22% of papers were left out. Despite a significant decrease in the number of authors, the percentage of women in the gender subset remains fairly stable at 18,52% compared to 17,95% for the entire data set.

5.1.3 Ethnicity

In this work, ethnicity denotes ’groups that share a common identity-based ancestry, language, or culture’ [87]. In order to include authors’ ethnicity in the analysis, the R-package WRU was used, that implements the algorithm discussed by Imai and Khanna [107]. This algorithm uses census data from the United States and predicts ethnicity based on last name and gender of an author.

However, the Genderize API is less accurate for researchers belonging to an ethnic minority, leaving out 10,09% of non-White authors after filtering for gender accuracy. Therefore, a separate dataset was created, that includes predictions of authors’ ethnicity.

For each individual researcher, the WRU algorithm returns five percentages (i.e. 0% Hispanic, 0% White, 0% Black, 99% Asian, 1% other). Only the ethnic groups for which the percentage is highest was retained, and any author whose ethnicity was predicted with an accuracy lower than 50% was filtered out. As a result of this filter, the dataset used contains 78 conferences, 105 988 papers and 76 749 authors. Despite the decrease in the number of authors, the percentage of researchers of colour in the ethnicity subset remains stable at 41,7% compared to 42,1% for the entire data set.

5.2 Results

5.2.1 Collaboration patterns in the co-authorship network

In order to analyse collaboration patterns in the co-authorship network, two different classes of metrics were used. Global network metrics allow to compare collaboration patterns between men and women, as well as between White researchers and researcher of colour. They were computed on subsets of the network, e.g. networks obtained by only considering men/women. Local network metrics focus instead on the performance and position of the individual researchers within the overall co-authorship network. In this case, metrics were computed on all nodes of the network.

Global network metrics include clustering coefficient (CC), average path length (APL) and diameter. The clustering coefficient measures network ’transitivity’ or the number of triangles in the network [108]. The average path length measures the average shorter path between any two nodes in the network. Finally, the diameter measures the longest shortest path between any two nodes in the network.

In this case, two separate men and women subnetworks were retrieved from the network by selecting only nodes representing men or women respectively and all the edges connecting them. Global network metrics were computed for both the women and women network, as well as for a null model that preserves the number of nodes and reshuffles edges for each subnetwork. The results for average path length and diameter were then normalised by comparing the original network to the null model. The clustering coefficient, instead, was normalised by the theoretical number of triangles in the network (between none or 0 and all or 1).

Local network metrics, instead, include betweenness centrality, closeness centrality, local clustering coefficient and degree. The betweenness centrality of a node measures the number of shortest paths between any two nodes in the network that go through that single node and indicates who occupies strategic positions in the network in terms of information exchange between researchers. The closeness centrality of a node measures the average shortest path between that single node and any other node in the network. The local clustering coefficient is a measure of a node's ego network density. Finally, the degree of a node indicates the number of direct collaborators of that node.

Results were separated by gender or ethnic groups and then they were ranked starting from top researchers.

5.2.1.1 Comparison of collaboration patterns for gender and ethnic groups

The differences in global network metrics are considered, according to gender, as shown in Table 5.1. The clustering coefficient is higher for the network of women than for men, meaning there is a higher transitivity in the the women subnetwork. Transitivity indicates that if researcher A collaborates with researcher B and researcher B collaborates with researcher C, it is more likely for researcher A to work with researcher C, forming a triangle. Since women are more likely to form triangles, this could explain why they are more likely to work with each other, which will be verified in the section dedicated to homophily. Furthermore, women might either share similar research interests or they might give the overlapping research interest a higher relevance which could result in more collaboration between women than between men.

Table 5.1: Global network metrics for women and men subnetwork.

	women	Men
Clustering Coefficient	0.27	0.24
Average Path Length	1.23	1.3
Diameter	1.82	1.99

The average path length is slightly higher for the network of men, meaning that the average shortest path between each couple of researchers in the network is slightly longer between men than between women. Men might be further removed from each other due to different research interests, or perhaps less collaboration between men in very disciplinary research areas [82]. Finally, the diameter of the women network is smaller than for the network of men, indicating that the longest short path between two women is still shorter. The small diameter cannot be attributed to the lower overall number of women in the network because the results were normalised. Rather, the women minority might have a more close knit network.

In the second stage the global network metrics were computed for ethnicity in Table 5.2. In order to compare majority and minority, all researchers of colour (Asian, Hispanic, Black and other) were combined into a subnetwork called ”people of colour” (PoC). The same method was used for normalising the results as the earlier comparison for gender.

Table 5.2: Global network metrics for PoC and White researchers subnetwork.

	PoC	White
Clustering Coefficient	0.2	0.25
Average Path Length	1.21	1.29
Diameter	3.04	2.13

First of all, the clustering coefficient is lower for people of colour which contrasts with the higher clustering coefficient for the women minority. One main difference is that there are more researchers of colour compared to women in Computer Science. So White researchers show higher transitivity even though they are in the majority. Perhaps White researchers are more likely to collaborate with each other than researchers of colour, thus displaying a higher ethnic homophily, which will be elaborated in the dedicated section.

The average path length for researchers of colour is slightly lower than for White researchers, yet the diameter is much larger for researchers of colour, which is remarkable because social networks usually show a high transitivity and a small diameter [108]. Researchers of colour might be more geographically dispersed and are therefore further removed from each other in the collaboration network. Another possible explanation could be that they work in many different areas of research that do not overlap.

In comparing the gender minority to the ethnic minority in computer science, the subnetwork of women behaves differently. The global network metrics for women confirms our expectations of higher transitivity and a closer collaboration, which contrasts with the characteristics of the ethnic minority. The differences in average path length are generally small however, and less indicative of different collaboration patterns between the majority and the minority.

5.2.1.2 Performance and position of computer science researchers

We consider the performance of researchers based on their position in the network and the structure of their ego network by computing local metrics for each individual researcher. The metrics used are betweenness, closeness, local clustering coefficient and degree. In this case, metrics were computed on all nodes of the network, that were then separated by gender or ethnic group and ranked starting from the top researchers.

Fig 5.1 shows the local network metrics for gender. Green/dark for men and yellow/light for women in the network. Men score higher for overall betweenness. However, women outperform men by a few percentage points in the top ranks, thus covering important positions in the flow of information. For closeness centrality, although women perform slightly better than average at the top ranks, men show a higher closeness overall, and thus dominate central positions in the network. This point will be further elaborated in the orbit analysis in section 5.2.1.3. The top ranks for clustering coefficient are male dominated, which stands in opposition to the global clustering coefficient where the women subnetwork scored higher. Meaning that when looking at the clustering of only women this is generally higher, yet in the overall network the top men outrank women in terms of transitivity. Finally, men outrank women in terms of degree for the first 150 researchers, meaning they have more direct connections with other researchers in the network. Only in the range between 150 and 250 women are slightly more connected than men.

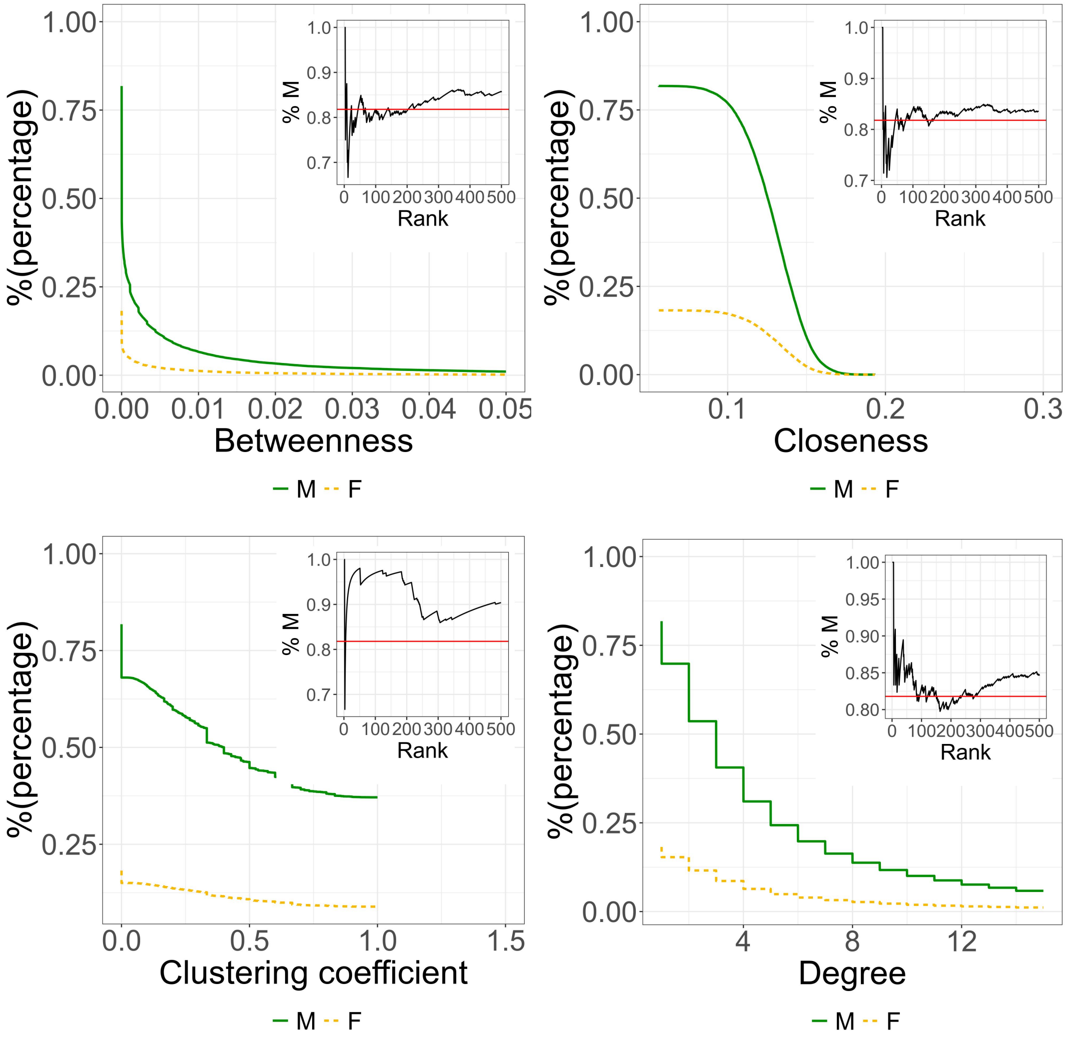


Figure 5.1: Local network metrics for gender. The x-axis shows the range of values for the metrics. The y-axis shows the percentage of nodes within a specific x-value. Dark green represents men, light yellow represents women. The insert shows the percentage of top men ranked from highest to lowest scoring for the specific metric. The red line represents the percentage of men in the entire network.

Fig 5.2, instead, shows the local network metrics for ethnic group. Light blue for White researchers, red for Asian, dark blue for Hispanic, orange for Black, and purple for other. All ethnic minorities are grouped as Researchers of colour, opposed to White researchers, in the insets. Overall researchers of colour have a higher betweenness, with the exception of a few White researchers in the top rank. Out of all researchers of colour, Asian researchers cover most strategic positions in the flow of information. The closeness centrality shows that researchers of colour perform even better this time. Overall, Asian researchers occupy most central positions in the network which will be further discussed in the orbit analysis section. The local clustering coefficient, instead, shows that White researchers, especially in the top ranks, are more clustered. The higher local clustering coefficient for White researchers, corresponds to a higher global clustering coefficient in the White researchers subnetwork. The degree, however, shows that researchers of colour and especially Asian researchers have more direct connections with others in the network. Despite the finding that Asian researchers cover most strategic and especially central positions in the network, White researchers have a higher transitivity. Perhaps White researchers are more clustered together and less open for collaborations with researchers of colour.

5.2.1.3 The role of individual researchers in collaborations

The role of researchers in a network can be assessed using graphlet analysis. Graphlets are small connected graphs [109], of which the size is determined by the number of nodes.

First, the orbit count of each node within the entire network was computed, which is the number of times a node assumes a specific role in any graphlet. Next, nodes were separated by gender or ethnic groups and the average number of times that a man/woman or White/PoC researcher appears in an orbit was calculated. Based on the following formula, a single value indicates whether an orbit is rather female or male dominated and to what extent: $\frac{O_i^M - O_i^F}{O_i^M + O_i^F}$. O_i^M represents the average count, over all nodes, of orbit number i for men and G_i^F represent the same quantity for women. If the result is negative, an orbit is female dominated, whereas a positive result indicates that an orbit is male dominated. The same formula was applied to determine whether an orbit is dominated by White researchers which returns a positive result, or dominated by researchers of colour which returns a negative result.

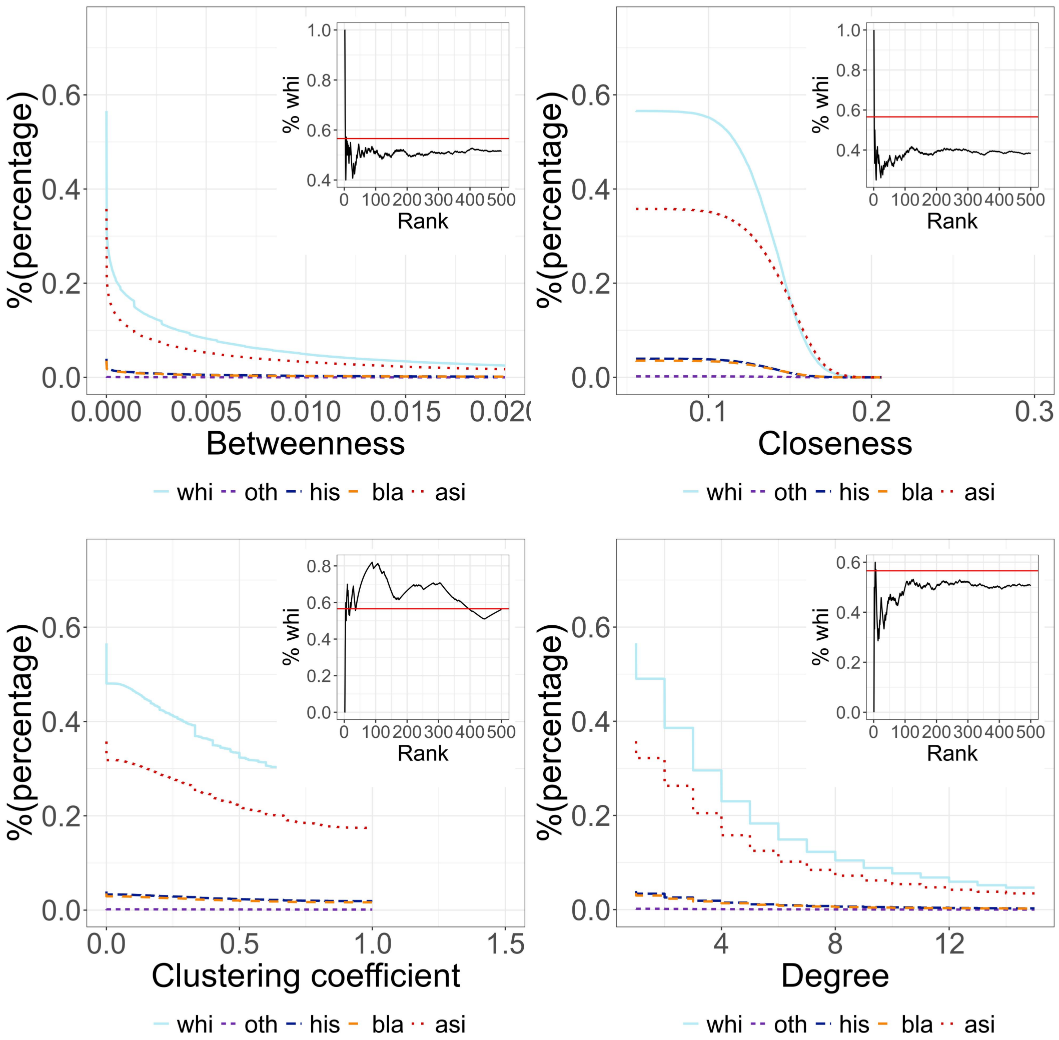


Figure 5.2: Local network metrics for ethnicity. The x-axis shows the range of values for the metrics. The y-axis shows the percentage of nodes within a specific x-value. Light blue represents White researchers, dark blue stands for Hispanic researchers, orange is used for Black researchers, and red stands for Asian researchers. The insert shows the percentage of top White researchers ranked from highest to lowest scoring for the specific metric. The red line represents the percentage of White researchers in the entire network.

For studying gender, the results were plotted in Figure 5.3 where colour labels male (green/dark) or female (yellow/light) dominated orbits and node size represents the likelihood that authors in these network positions are men or women. All orbits with a central role (in graphlet G_1, G_3, G_4 and G_6) are male dominated (orbit 2, 5, 7 and 11). Orbit 11 in graphlet G_6 is of particular interest. This specific orbit identifies a brokerage position that is covered mostly by men, mediating between a pair of women that are connected to each other (orbit 10) and an individual woman (orbit 9). Furthermore, the only female dominated orbits (in graphlet G_1, G_3, G_4 and G_6) are all relegated to the peripheral positions in the network (orbits 1, 4, 6, 9, and 10).

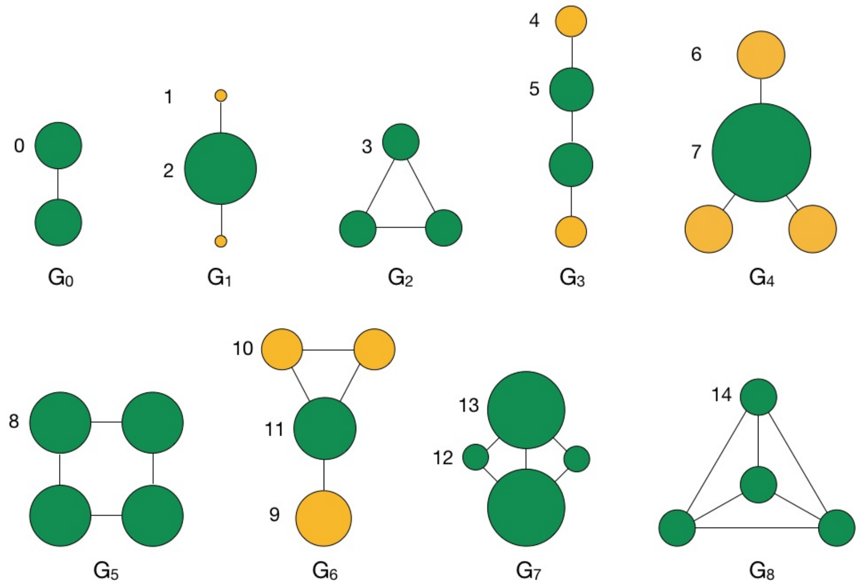


Figure 5.3: Graphlets up to 4 nodes with their relative orbits. Green/dark represents male dominated orbit positions, yellow/light represents female dominated orbit positions. Size represents the probability that authors in these network positions are men or women.

The same analysis was performed for White researchers and researchers of colour. The results are shown in Figure 5.4 where colour labels White dominated orbits (blue/-light) or orbits dominated by researchers of colour (purple/dark). Node size represents the likelihood that authors in these network positions are White or people of colour.

The first main finding was that not a single orbit is dominated by White researchers. Secondly, a similar majority/minority patterns can be noticed in the ethnicity orbit analysis as for gender. Meaning that all orbits with a central role (in graphlet G_1, G_3, G_4, G_6) are less dominated by researchers of colour. The brokerage position of orbit 11 in graphlet G_6 is also slightly more often covered by White researchers. Less outspoken perhaps, yet still visible, is the fact that peripheral positions in the network (orbits 1, 4, 6, 9, and 10) are more often dominated by researchers of colour. The overall dominance of researchers of colour in the orbit analysis corresponds to the dominance of researchers of colour for all but one of the global network metrics. Only in terms of transitivity do White researchers outperform researchers of colour for the global network metrics, which translates to almost equally distributed orbit positions (3, 14) where transitivity is apparent.

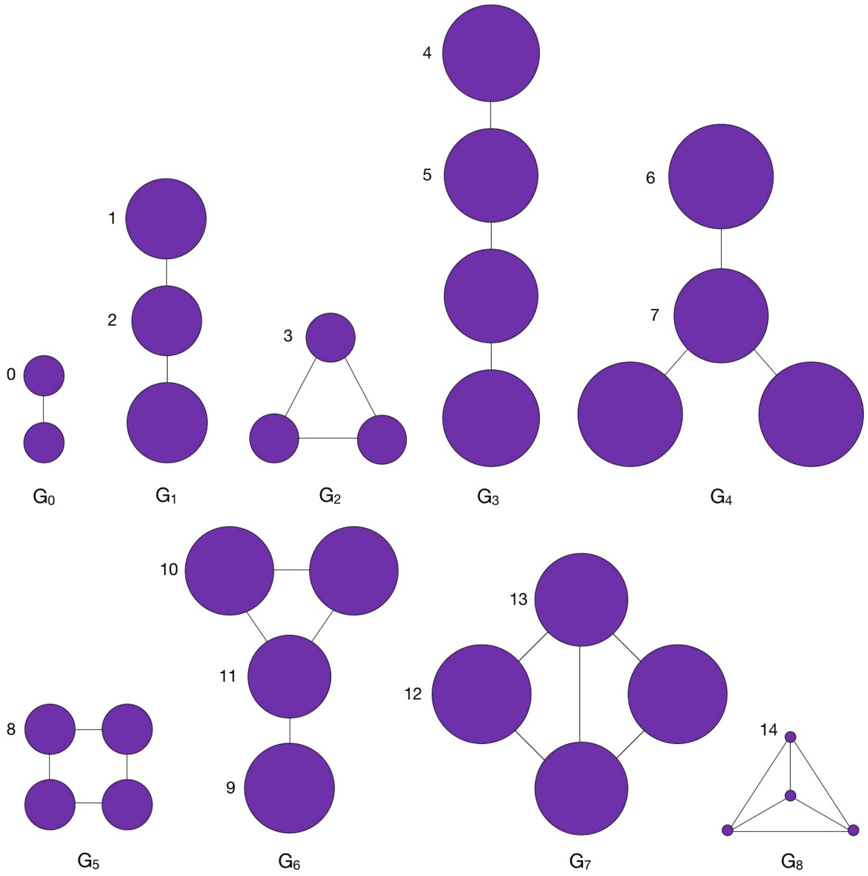


Figure 5.4: Graphlets up to 4 nodes with their relative orbits. Dark/purple represents researchers of colour dominated orbit positions, and light/blue represents White dominated orbit positions, (however, no orbits are dominated by White researchers). Size represents the probability that authors in these network positions are White or people of colour.

5.2.2 Evolution of the co-authorship network

In order to analyse collaboration patterns over time, a temporal version of the co-authorship network previously discussed was built. Given a sliding time window of length

d and a time step s, several subnetworks were produced. For example, if you consider the interval [1960, 2015], a time window of five years and a time step of one year, the networks corresponding to each time slice will include all nodes that are active in the interval [1960, 1964], [1961, 1965], ..., [2011, 2015] and their relative connections. A node is considered active in a certain interval if a paper by that author was published within the year interval.

5.2.2.1 Do researchers flock together based on gender or ethnicity?

Homophily is the principle for which a contact between similar people occurs at a higher rate than among dissimilar people, for example that researcher of the same sex/ethnicity are more likely to collaborate. Social groups, organisations and any positions in social systems all create contexts in which homophilous relations form, which in turn could result in the formation of niches [54].

Fig 5.5 shows the evolution of gender and ethnic homophily over time in the co-authorship network. A sliding window of ten years and a step of one year have been used. The result was normalised according to the size of the relative group (men and women, White researchers and researchers of colour). Therefore, homophily will be greater than the proportion of the group if there is intra-group homophily (people from the same group are more likely to co-author a paper), or lower if there is inter-group homophily (people in the same group are less likely to co-author a paper).

The results show an overall intra-group homophily for both gender and ethnicity. In terms of gender, the proportion of men decreases over time. Male homophily decreases, following the same trend, but remains constant with respect to the group proportion, while female homophily increases starting from the 1980s and decreases again around 1995. For ethnicity, there is a similar trend where the proportion of White researchers decreases over time. White homophily decreases over time but it grows when compared to the group proportion, whereas homophily within researchers of colour increases over time at a faster rate than the group size.

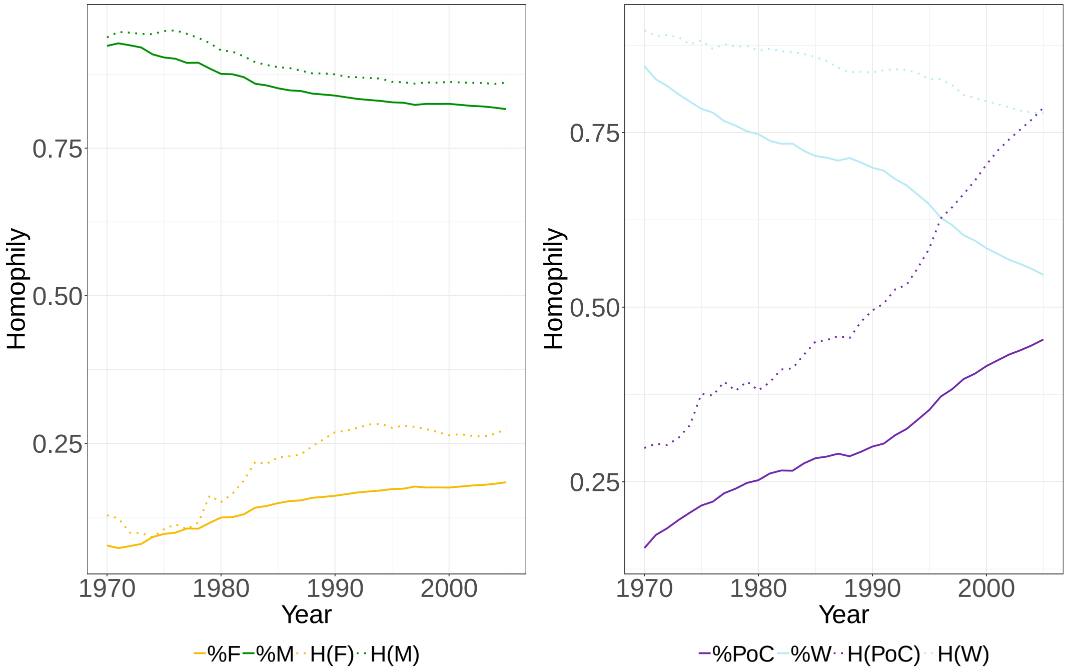


Figure 5.5: Homophily over time. Window size $d = 10$ and time step $s = 1$. Year on the x-axes refers to the interval [$\bar{x}$, $\bar{x}$+d]. Colour indicates different groups. Solid lines represent the proportion of a group, while dotted lines represent group's homophily

5.2.2.2 The changing position of computer science researchers over time

For the static network, the performance of researchers based on their position in the network, as well as the structure of their ego network, were analysed by computing local metrics on each node of the network. In order to analyse these characteristics over time, a temporal network was built, using a sliding window of ten years and a one year step. Betweenness, closeness, local clustering coefficient and degree were computed for each node over time. Since the composition of the group of researchers changes for every time window, the ratio between men and women or between White researchers and researchers of colour was normalized to zero. For example, the result is zero when the ratio between men and women in the top 100 researchers is exactly the same as for the overall network. It does not mean that the ratio between men and women is 50/50.

In Figure 5.6 the local network metrics for gender are shown. For betweenness and closeness centrality, as well as degree, it was already established that men (green/dark) cover most of the important and central positions in the flow of information and they have most direct connections, but that women (yellow/light) outperform men in the top ranks. The evolution over time shows, however, that women only occupy the first top ranks in the middle of the 1980s and in the middle or at the end of the 1990s, and the top 100 ranks over the last twenty years. The clustering coefficient shows that women in the top ranks during the 1970s were decreasing in the ranking as men entered the field in the 1980s and again at the start of the 1990s. Since 1995 the overall transitivity for the top ranks in terms of clustering coefficient is clearly higher for men.

The local network metrics for ethnicity can be found in Figure 5.7. A similar pattern in terms of betweenness, closeness, and degree arises for ethnicity where overall researchers of colour make up the majority of the top ranks. They cover most of the important and central positions in the flow of information and have more direct connections than their White counterparts. Until the mid 1990s, researchers of colour dominate the very top positions, and in the mid 1980s researchers of colour also make up the top 250 researchers. However, White researchers began to cover the central positions in the network (closeness centrality) since the beginning of the 2000s. The clustering coefficient shows that the transitivity of White researchers in the top ranks since the 1980s has decreased significantly over time. In the last fifteen years, researchers of colour have become prominent in the top 50 ranks for transitivity.

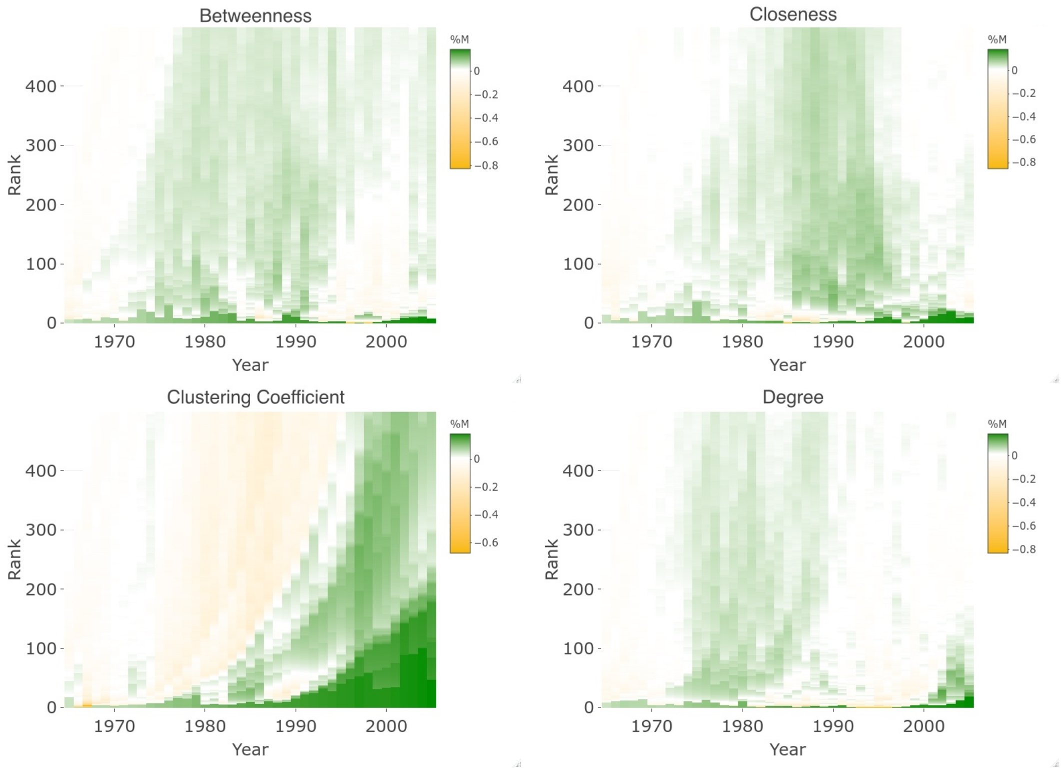


Figure 5.6: Local network metrics over time for gender. The x-axis shows the time interval of ten years starting at year $\bar{x}$, while the y-axes shows the top $\bar{y}$ researchers, ranked from highest to lowest score for a specific metric. Colour represents the percentage of men on the top $\bar{y}$ ranks in time interval $\bar{x}$. Dark green (positive values) indicates an over-representation of men, whereas light yellow (negative values) indicates an over-representation of women. White (values close to zero) indicates that the percentage of men and women in the top $\bar{y}$ ranks reflects the percentage in the whole network

5.2.2.3 The evolving roles of researchers in collaborations

For the static network, the role of individual researchers in the collaboration network was already analyzed through orbit analysis. To analyze their evolution in time, a temporal network was constructed and the orbit counts were computed for each ten year time window starting in 1970, as a metric for the evolving roles of researchers. The first five years were left out because the data was too sparse to draw any significant conclusions.

We previously established that in terms of gender balance all orbits with a central role (2, 5, 7, 11) are male dominated. As Figure 5.8 shows, the peak for men in central orbit positions was between 1970 and the end of the 1980s. Yet even after the 1980s, central

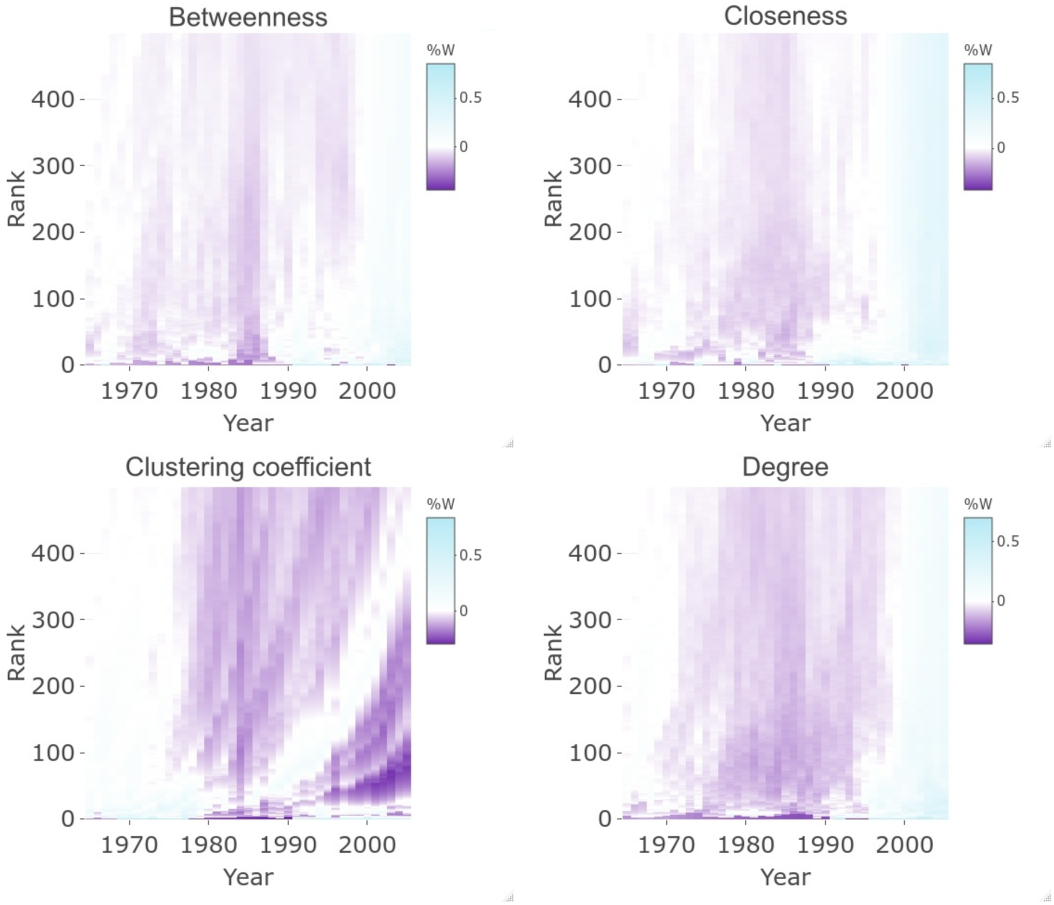


Figure 5.7: Local network metrics over time for ethnicity. The x-axis shows the time interval of ten years starting at year $\bar{x}$, while the y-axes shows the top $\bar{y}$ researchers, ranked from highest to lowest score for a specific metric. Colour represents the percentage of White researchers on the top $\bar{y}$ ranks in a time interval $\bar{x}$. Light blue (positive values) indicates an over-representation of White researchers, whereas dark purple (negative values) indicates an over-representation of researchers of colour, whereas . White (values close to zero) indicates that the percentage of White researchers and researchers of colour in the top $\bar{y}$ ranks reflects the percentage in the whole network

positions were rarely female dominated, with some minor exceptions in the middle of the 1990s for orbit 7 and the brokerage position in orbit 11. Overall it was already found that women only dominate peripheral positions (orbits 1, 4, 6, 9, and 10). These results show a reverse trend where peripheral positions become female dominated from the mid-1990s onward. In conclusion, the gender imbalance was certainly more pronounced in the past in terms of what roles individual researchers cover in the co-authorship network, than towards the end of the 2000s.

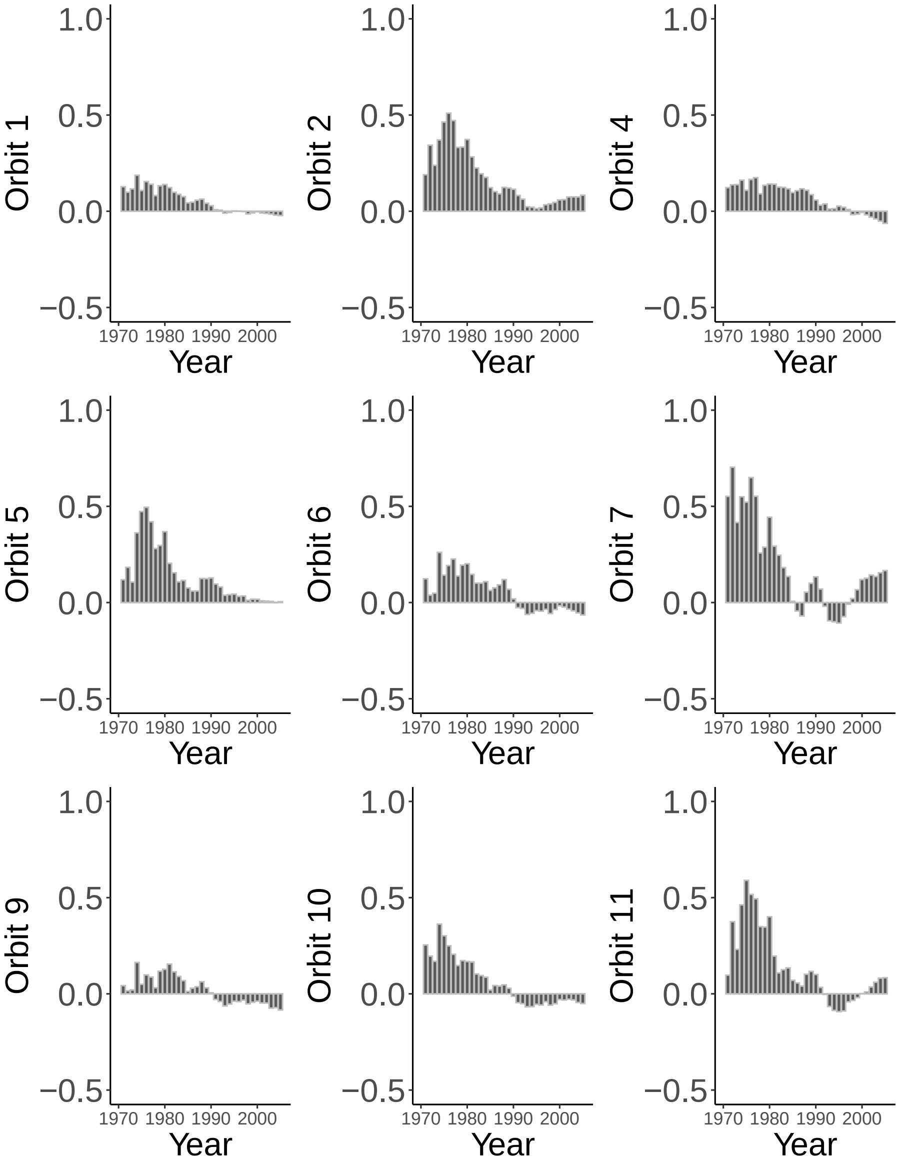


Figure 5.8: Each graph presents the evolution of the composition of a specific orbit between the 1970s and 2015. The x-axis shows the time interval of ten years starting at year $\bar{x}$. The y-axis displays whether an orbit is on average more dominated by either men (positive values) or women (negative values) researchers.

Although the orbit analysis for ethnicity on the static network showed that researchers of colour dominated all orbit positions, Figure 5.9 shows that this is a recent phenomenon. Until the end of the 1990s in fact, just about every single orbit position was dominated by

White researchers. Furthermore, the overall smaller prevalence of researchers of colours in central orbits (2, 5, 7, 11) is due to the dominance of White researchers up until the end of the 1990s, rather than the current prevalence of researchers of colour (mostly Asian researchers).

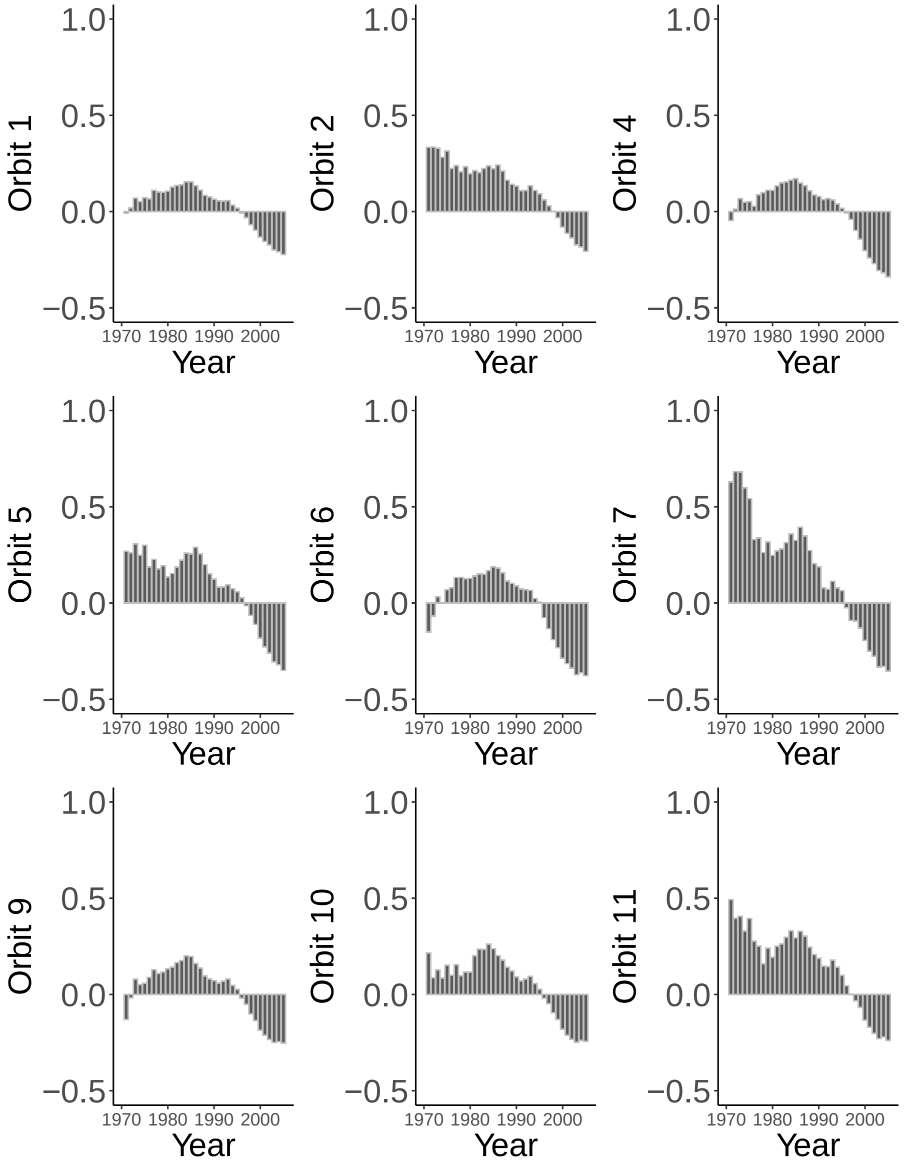


Figure 5.9: Each graph presents the evolution of the composition of a specific orbit between the 1970s and 2015. The x-axis shows the time interval of ten years starting at year $\bar{x}$. The y-axis displays whether an orbit is on average more dominated by either White researchers (positive values) or researchers of colour (negative values).

5.3 Limitations of the study

This work shows the gender and ethnic differences in collaboration patters in the Computer Science community. Given the large number of authors and limited personal information available, an automatic method to classify authors' gender and ethnicity was necessary. In order to use such tool, some generalizations/simplifications had to be made and the use of categories was necessary. We assumed that gender is binary instead of a spectrum, and we did not consider authors whose gender was classified as unknown, even though distributed residual categories, for example "other" fields, can add another dimension to the data, preserve complexity and indicate uncertainty [110]. The same applies for ethnicity. We assumed that an author can only belong to one ethnic group, and did not consider mixed ethnic groups.

It is important to recognize that these classifications are not always fair and can have consequences. In fact, they can produce advantage or suffering whether a person is set, within the classification system, in a position of power or not [110]. It is also important to recognize the limitations of automated gender prediction. Algorithms are not free of bias and the choice of one algorithm over the other can introduce a bias already.

The use of automated methods and categories are intended to provide a simplified model that can be analyzed and used to understand a more complex reality.

5.4 Conclusion

This work shows the gender and ethnic gap in computer science research by looking at different collaboration patterns in the co-authorship network. Network analysis metrics were used to quantify researchers' position in academia. It was found that women, overall, score lower than men in terms of network connections and are more close-knit. However, they do perform better at the top ranks and some women do cover central positions in the co-authorship network. Asian researchers make up most of the researchers of colour, and they cover strategic and central positions in collaborations, outperforming White researchers. White researchers, in the same way as women, are more close-knit. Looking at these trends over time, it is noticeable that the prevalence of one or the other gender or ethnicity has evolved. In fact, women perform better at the top ranks only during specific periods, such as in the middle of the 1980s and since the end of the 1990s. The trend for ethnicity instead inverted from researchers of colour covering central and strategic positions until the mid-1990s and becoming more close-knit in the last 15 years.

Using orbit analysis, it was found that men mostly cover central positions in the co-authorship network, while women are relegated to the periphery. Researchers of colour, instead, cover most positions due to their higher number of direct connections. In general, gender differences have narrowed over time, while we can observe a complete inversion of the trend for ethnicity. In fact, up to the 1990s White researchers were dominating most orbits.

Finally, minority group (women and researchers of colour) are expanding over time, with the intra-group homophily increasing even faster.

CHAPTER VI

INTERPLAY BETWEEN SUCCESS AND PATTERNS OF HUMAN COLLABORATION: CASE STUDY OF A THAI RESEARCH INSTITUTE

This chapter is partially based on the published article “Interplay between success and patterns of human collaboration: case study of a Thai research institute” [111].

Networks are used to model different systems such as biological ones (Jeong et al. [36]), the world wide web (Réka, Hawoong and Barabási [35]), organizations and societies. Social Network Analysis is a truly interdisciplinary domain that has gained traction due to the recent access to large-scale datasets (“Big Data”) available online. A social network can be described as a collection of actors that are connected to each other if they form some sort of relationship [47]. A collaboration network is a particular social network where nodes represent individuals belonging to an institution/organization/company and edges represent collaboration and/or interaction between individuals [112]. Networks of collaboration are notoriously complex and the mechanisms underlying their evolution, although of high interest, are still not fully understood. In particular, collaboration networks can be used to model the interactions between scientists and analyze the circumstances that lead to successful research.

6.1 Science of collaboration

Scientific success and productivity have a skewed distribution [16]. Elmacioglu and Lee [97] have shown how a small fraction of authors publish a large number of papers. Seglen, similarly, has shown that a small portion of articles collect most of the citations [16]. Newman and Girvan [98] uncovered that the distance between scientists in a collaboration network is typically small (i.e. small world property) and that, for most scientists, all paths between them and other scientists go through only one or two of their collaborators (the so-called funneling effect). In the work of Backstrom et al. [101], attention was drawn to the evolution of communities in collaboration networks and the structural features that influence the decision of individuals to join a community. Longitudinal analyses were also performed by Huang et al. [102] and Bird et al. [102] to find the structural differences between topical areas, identified as communities.

The factors that define a successful collaboration are various and of different nature. Mattessich and Monsey, in their book [113], review the existing literature on which factors influence the success of collaborations, as well as what measures can be taken to enhance fruitful collaboration. They define collaboration as “a mutually beneficial and well defined relationship shared into by two or more individuals or organisations to achieve common goals”. They classify these factors in the following categories: environment, membership, structure, communication, process and resources. Amongst these many factors, the ones related to communication are worth mentioning. They found that established informal and formal communication links play an important role in successful collaborations: members that establish inter-personal relationships produce better results when working on a common project. Also, members that engage in open and frequent communication favour the transmission of information within and outside the group.

One of the biggest challenges is the one to find any links between patterns of collaboration and scientific success. Authors, with the increasing pressure to publish more, tend to seek for more collaborations [97]. Collaborators can have a large effect on a researcher's career and choosing the right collaborators can have long-term implications on access to knowledge and resources. Borjas and Doran [114] argue that spillover exists in three dimensions: idea (working on the same topic), geographic (working in the same department/university/region) and collaboration (co-authorship). They put the spillover hypothesis to the test, which states that emigration of researchers in any of these dimensions would lead to a reduced productivity for the researchers left. They found that spillover does not affect the average researcher, but it will affect researchers who lose a regular coauthor. Petersen [115] shows that researcher's collaboration patterns consist of high turnover collaborations, identified in weak ties, as well as steady and frequent collaborations with "life partners", identified in super ties. He presents the "apostole effect", that illustrates the advantage of strong and committed relationships. A longitudinal analysis was reported by Abramo et al. [116] to find a causal effect between collaboration patterns and performance. They found that researchers who moved higher in rank tend to have fewer intramural collaborations, while favoring international ones, thereby leading to publications of higher impact. Feng and Kirkley [117] analyzed the link between researchers' neighborhood structure and academic performance. They found that researchers who collaborate with many teams and work on several projects have a longer career and are highly performing. Cross, Borgatti, and Parker [118] found that even informal networks, such as friendship, contribute positively to job satisfaction and performance. They argue that, even if these sort of networks cannot be directly controlled by management, they can still be affected by factors such as hierarchical levels, office location, project staffing and so on.

Petersen et al. [119] underline the urgency of defining new performance measures for individuals and groups. Conventional metrics, based on number of publications and number of citations, may not be sufficient to provide a deep insight into the factors driving scientific success. Seglen [16], for example, argues that the skewed distribution of authors' citations and the different citation practices in each field of research make citation count not suitable to evaluate researchers' success. Instead, network analysis techniques based on centrality measures could be used to shed a new light on some mechanisms of success. Centrality metrics were first used by Bavelas [41, 42] who linked these metrics to team performance and productivity. Uddin et al. [120] used both correlations and regression methods to link centrality measures and performance, showing that scientists that cover central positions in the network (high betweenness centrality) and have many collaborators (high degree centrality) are also highly cited. Similarly, Sarigöl et al. [121] showed that centrality measures (degree centrality, eigenvector centrality, betweenness centrality and k-core centrality) of authors at the time of publication are good predictors of their citation count in the following five years.

In recent years, interest moved from individuals to teams. It has become more and more apparent that the most successful research is carried out by teams rather than single researchers [122]. It has been found that, over the past 50 years, teams increasingly dominate solo authors in the knowledge production. In addition, teams are more likely to produce high impact research in academia as well as in the private sector [123]. Uzzi et al. [124] also found that teams are 37.7% more likely than solo authors to bring novelty into established knowledge domains, and that papers of this type are twice as likely to be highly cited. Petersen et al. [119] show that scientific productivity is related to researchers' visibility and team efficiency. In fact, teams can produce higher impact output due to the larger number of coauthors involved, that will be able to introduce their work to more peers.

For all these reasons, researchers have become interested in studying the dynamics of social groups, which consists of series of changing events such as formation and dissolution of teams. Guimera et al. [125] proposed a method for group evolution discovery, based on a similarity measure between groups, and showed that the most successful teams are the ones that have a large core of established members who actively seek for new collaborations. Palla et al. [126] proposed another technique, based on the Clique Percolation Method, to investigate the evolution of groups over time. It was found that large and small groups behave differently. Specifically, large groups have a higher lifespan when turnover is high, meaning that there is a constant flow of newcomers, while small groups are better off when their composition remains unchanged over time. On the other hand, Kenna and Berch [127] propose the notion of critical mass, for which research quality increase with group size only up to a maximum size referred to as critical mass. Goa et al. [128] proposed a different method, based on central nodes identification, to study patent classes of similar technologies. Finally, Reagans and Zuckerman [129] put to the test two hypotheses: the closure view of social capital and the structural hole view on social capital. The closure view of social capital states that teams that experience more frequent communication among their members (higher density) can achieve higher productivity [130]. The structural holes view on social capital, instead, states that teams that experience more frequent communication among members with different attributes (more heterogeneous) achieve a higher level of productivity [131]. They found both hypotheses to hold true and, particularly, that team density is more advantageous for heterogeneous teams.

In this study, we use a dataset from the National Electronics and Computer Technology Center (NECTEC) in Thailand, where researchers collaborate on different projects and team up to produce a range of artifacts (intellectual properties, prototypes and scientific articles). For each artifact, a score that measures quality of research is available and shared between the researchers that contributed to its creation, according to their percentage of contribution. We build a collaboration network where researchers are connected if they worked together on one or more artifacts.

6.2 Methods

6.2.1 Dataset

The dataset includes information about projects carried on at the NECTEC institute for a nine year period, from 2009-10-03 to 2018-07-26, for a total of 553 projects. Within each project, researchers collaborate to produce three different types of artifacts: scientific articles, prototypes and intellectual properties (IP). The dataset contains 1202 records for articles, 459 for prototypes and 631 for IP, for a total of 8531 collaborations. Time information is also included in the dataset, as filing date, for each artifact. The dataset is publicly available at `https://github.com/apivadee/research-collaboration`.

6.2.2 Building the network

The constructed collaboration network represents researchers as nodes, that are connected to each other if they have collaborated to produce one or more artifacts. The NECTEC researchers network includes 740 nodes/researchers and 5298 edges/unique collaborations and is undirected. Nodes are assigned the following attributes:

- `n.artifacts`: number of artifacts produced by a researcher.
- `n.projects`: number of projects a researcher has participated in.
- `score`: total IC score assigned to all artifacts produced by a researcher, weighted by their contribution.
- `percent.contrib`: average percentage of contribution of a researcher for all artifacts they have worked on.
- `start.career`: start of a researcher's career, i.e the filing date of the first artifact produced by a researcher.
- `end.career` : end of a researcher's career, i.e. the filing date of the last artifact produced by a researcher.

Edges, instead, have the following attributes:

- `n.artifacts`: number of artifacts that two researchers have collaborated on. This quantity is normalized by the total number of artifacts produced by both researchers.
- `n.projects`: number of common projects that two researchers have participated in. This quantity is normalized by the total number of projects that the two researchers participated in.
- `contribution.symmetry`: indicates how much, on average in a collaboration, one researcher contributes to the production of their artifacts compared to the other. It ranges from -1 to 1 in case only one researcher contributes to all the work, and 0 in case of equal contribution.
- `start.date`: start date of a collaboration, i.e. the filing date of the first artifact produced by two researchers.
- `end.date`: end date of a collaboration, i.e. the filing date of the last artifact produced by two researchers.

A dynamic version of the network previously discussed is built. Given a sliding time window of length d and a time step s, the static network is "sliced" and a subnetwork is produced for each time slice. For example, if you consider the time interval [01/01/2008, 01/01/2018], a time window of 3 years and a time step of 1 year, the subnetworks will include all nodes that are active in the interval [01/01/2008, 01/01/2011], [01/01/2009, 01/01/2012], ..., [01/01/2015, 01/01/2018]. A node is considered active in a certain time slice if its starting time or ending time falls within the interval. The dynamic version of the NECTEC researchers network, in particular, consist of seventeen time slices.

6.2.3 Performance metrics

The Intellectual Capital (IC) score, defined by NSTDA, is used to define the capital level of each R&D output within the NECTEC. It is assigned to each artifact and divided among researchers that worked on it, depending on their percentage of contribution. We have defined two different performance metrics for individual researchers: productivity metrics are related to the amount of work of a researcher, while quality metrics are related to the impact that a certain work has. As productivity metrics, we used number of artifacts produced by a researcher, number of projects that a researchers have joined, career length and number of collaborations. As quality metrics, we used the average IC score.

6.2.4 Network metrics

Global network metrics are metrics computed over the entire network. Among these metrics there are diameter, average path length and clustering coefficient. The diameter is the longest shortest path between any two nodes in the network, while the average path length is the average shortest path between any two nodes in the network. These two metrics represent how easily information can travel through the network. Clustering coefficient is computed as the ratio of the number of triangles and the connected triples in the network.

Local network metrics are metrics computed on single nodes. Among these metrics there are local transitivity and centrality metrics such as degree, betweenness and closeness. The local clustering coefficient is the ratio of the triangles connected to a node and the triples centered on the node. This metric is related to the concept of transitivity. In the collaboration network, if researcher A is connected to researcher B and researcher B is connected to researcher C, what is the probability that researcher A is connected to researcher C? The degree centrality consists in the number of direct connections of the node and represents the number of collaborators of a researchers. The betweenness centrality of a node is computed as the number of shortest paths between any couple of nodes in the network that pass through the node. It represents how critical is the position of a researcher in the network for the transmission of information. For example, nodes with high betweenness often serve as bridges between different communities. The closeness centrality is computed as the average shortest path between the node and any other node in the network. It represents how far information has to travel from a node to reach the entire network.

6.2.5 Team dynamics

Team dynamics or social groups dynamics is the analysis of team evolution. There has been different definition of groups [132, 133], coming from different disciplines. In Social Network Analysis, a group is defined as a set of actors that are highly connected to each other, when compared to the rest of the network [47]. This is a criterion for group existence rather than a proper definition of group. Therefore, depending on the scenario and needs, different definitions can be considered [134]. The social groups evolution problem can be decomposed into different steps:

- *Temporal network creation*: a temporal network is created by slicing the static network and extracting a network for each time slice. Networks can be generated by selecting the subset of nodes or edges active at a certain time [126] or within a time window. Time windows can be distinct [126] or overlapping [128] and their size can be constant or adapted to include a fixed number of nodes or edges.
- *Group identification*: for each network, groups are identified. Groups can be disjoint [128, 134] or overlapping [126, 134]. Also, networks can be treated independently [126,134] or the group identification at time t can take into account the groups found at time $t - 1$ [128].
- *Group tracking*: the evolution of groups is tracked by matching groups at consecutive time slices. Methods are based on similarity measures between groups [134], central nodes identification [128] or other methods that are suited to a specific community detection algorithm [126].

- *Event identification*: each matching between groups at consecutive time slices is associated to an event. Palla, Barabási and Vicsek [126], as well as Bródka, Saganowski and Kazienko [134] proposed seven different event types:
  - *continuing*: a group continues its existence when two groups at consecutive times are identical or almost;
  - *growing*: a group at time t grows when few nodes join the group, making its size slightly larger at time $t + 1$;
  - *shrinking*: a group at time t shrinks when few nodes leave the group, making its size slightly smaller at time $t + 1$;
  - *merging*: several groups at time t cease to exist and merge to form a new group at time $t + 1$;
  - *splitting*: a group at time t ceases to exist and splits to form several new groups at time $t + 1$;
  - *forming*: a new group forms when a group, that did not exist in time window t, appears in time window $t + 1$;
  - *dissolving*: an existing group dissolves when a group, that existed in time window t, does not exist anymore in time window $t + 1$;

Two definitions of groups are considered. The first one is related to the projects: a group is identified by all the people working on the same project. In this case groups can overlap, since one person can work on several projects at the same time. The second definition is related to the outcome of a community detection algorithm. Using a community detection algorithm whose output is a disjoint set of groups will give a significantly different outcome when compared to the first one, i.e. a single group will contain researchers working on two different projects, instead of two overlapping groups.

For this work, we used a sliding time window with length d and step s such that a temporal network with 17 time slices is generated. As group identification method, we decided to use MemLPA [59], a community detection method based on label propagation. In this case, communities are disjoint and the outcome of the algorithm at time t depends on the groups found at time $t-1$. As group tracking and event identification methods, we used the ones described by Bródka, Saganowski and Kazienko [134], that are based on a similarity measure called inclusion. According to the experimental analysis in their work, the optimal values for α and β are between 0.5 and 1. Using two different values for these parameters would favour the identification of growing/merging events over shrinking/splitting events or vice versa. Therefore, the parameters we used in our experimental analysis were set to $\alpha = 0.5$ and $\beta = 0.5$, and communities with size five or lower were omitted.

6.2.6 Team metrics

Two different type of metrics have been defined for teams: team performance metrics and team structure metrics. Team performance metrics include the number of artifacts produced by all team member (productivity metric) and the average IC score of all team member (quality metric). These quantities are normalized by team size. Team structure metric include team size, autocorrelation, density and ratio between outer and inner connections. Autocorrelation measures teams' turnover, hence the rate at which researchers join and leave a team. It is computed as the intersection between a team at time t and time $t+1$. Low/high autocorrelation indicates steady/dynamic teams. Density indicates the density of connection within a team. It is computed as the number of connections between team members and the total number of possible connections. Low/high density indicates sparse/close-knit teams. The ratio between outer and inner connections measures in what extend team members engage in collaboration with members of other teams. It is computed as the number of edges between team members and non-team members, divided by the sum of team members' degree. A low/high ratio indicates the tendency of teams to be closed/open. For each team, these metrics are computed for each time slice of their lifespan.

6.3 Results

We analyzed the distribution for some of the attributes for the NECTEC researchers network. For node attributes we considered number of artifacts, number of projects, Intellectual Capital (IC) Score and career length, while for edge attributes we considered number of artifacts, number of projects, contribution symmetry and length of collaboration. Results are shown in Figure 6.1. As can be noticed, few researchers produce a large number of artifacts, while most researchers produce only a few. Number of projects and IC score are distributed in a similar manner, where few researchers are able to carry out many projects at the same time, while having a high IC score. Similar results can be observed for edge attributes. These results reflect the skewed distribution of scientific productivity. Regarding the career lengths, most researchers are active for no more than 75% of the entire time frame, with very few exceptions.

We then analysed the evolution of the researchers network over time by computing some global metrics such as number of nodes, number of edges, global transitivity, diameter and average path length for each time slice. Figure 6.2 shows the results. In general, the first time slices contain very few nodes. This is because of a few projects starting off early. After that, more researchers become active for few time slices, but time slices 8-12 show again lower activity. Higher diameter and average path length seem to be indicators of a less active and less dense network, while transitivity reaches its lowest value.

6.3.1 Analysis of researchers' collaboration patterns and performance

We computed correlations between a combination of performance metrics and local network metrics. We used number of artifacts, number of projects and career length as

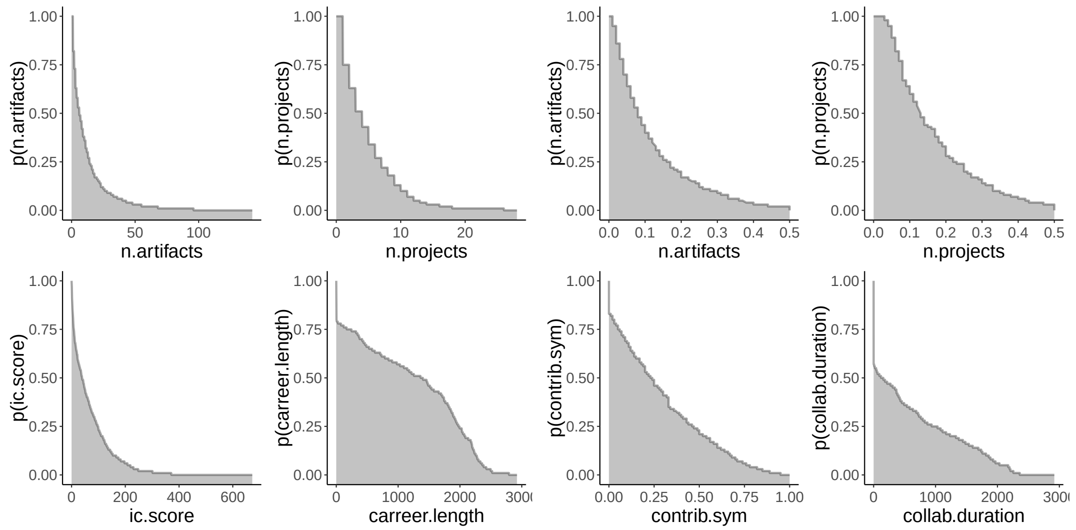


Figure 6.1: Node (left) and edge (right) attribute distributions for the NECTEC researchers network. The *x*-axis shows the range of an attribute value on a linear scale (see Methods section for definition of these attributes). The *y*-axis shows the probability that a node has that attribute value or greater. The attributes *career.length* and *collab.duration* are measured in days.

productivity metrics, as well as average IC score as quality metric (see Methods section). For local network metrics, we use betweenness, closeness, degree and transitivity. Figure 6.3 (left) shows the correlation matrix for the NECTEC researchers network. Number of artifacts, number of projects and career length correlate positively with degree, since researchers that are more productive or have a longer career also have more opportunity to engage in new works and collaborations. Number of artifacts, number of projects and career length also correlate positively with betweenness and closeness centrality. Therefore researchers that cover more central positions in the network appear to be more productive. Finally, number of artifacts, number of projects and career length correlate negatively with transitivity. The average IC score shows a different trend. It is not correlated with betweenness and closeness centrality, and it correlates negatively with degree and transitivity. This result suggests that researchers that have fewer collaborations at the same time produce higher quality work. As it can be noticed, centrality metrics show similar correlations with other metrics. Therefore, we decided to disentangle the link between them by normalising

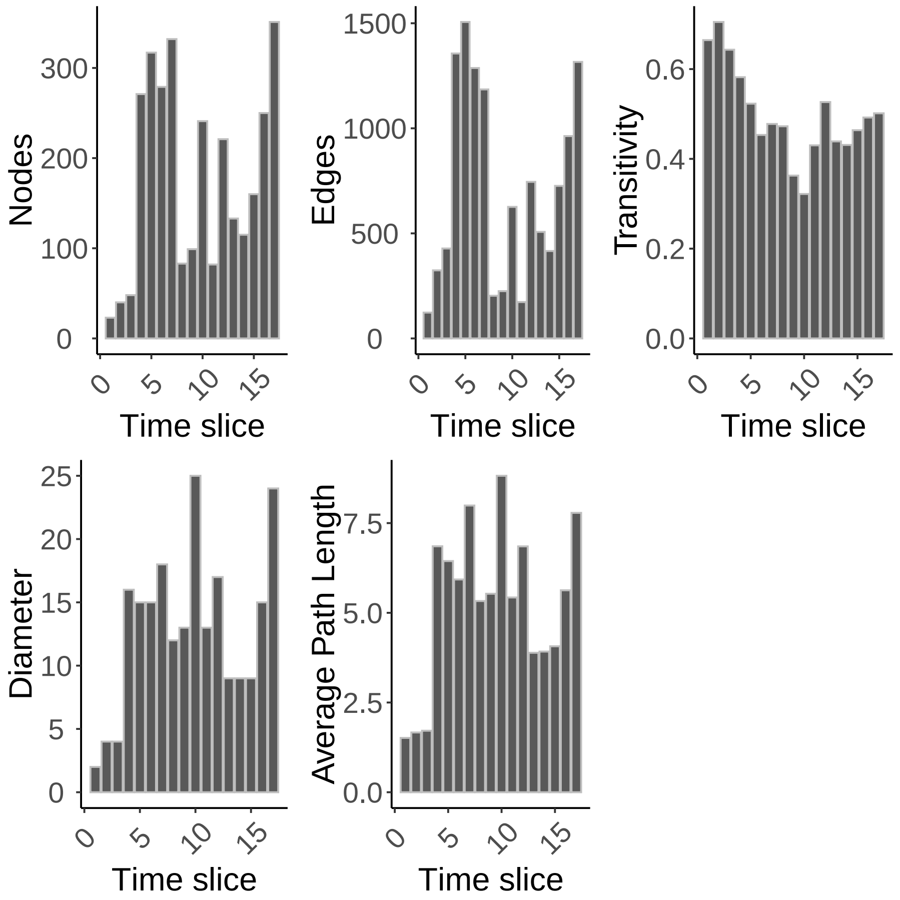


Figure 6.2: Global network metrics for the NECTEC researchers network.

betweenness and closeness by degree centrality. There, figure 6.3 (right) shows the same correlation analysis, where betweenness and closeness centrality are normalized. Results are slightly different. Betweenness centrality is still positively correlated with number of artifacts, number of projects and career length, even when normalized by degree. Closeness centrality, instead, is negatively correlated to the same quantities. This shows that high degree centrality alone cannot explain high productivity metrics such as number of artifacts/projects and long career length, and betweenness still plays an important role. Not only is the number of connections relevant, but also the position of a researcher in the network.

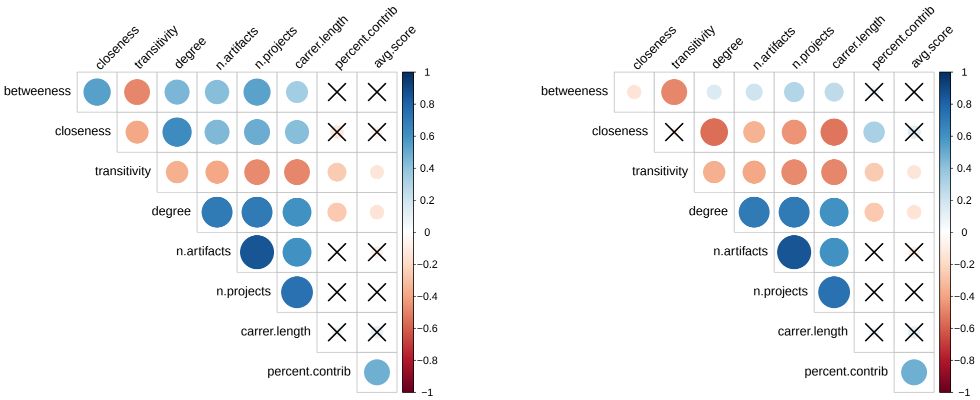


Figure 6.3: Correlation matrix for the NECTEC researchers network. Blue indicates positive correlation and red indicates negative correlation. Color intensity indicates the strength of correlation. Insignificant coefficients, according to p-value $p = 0.001$, are marked with a cross. On the right, betweenness and closeness centrality have been normalized by degree centrality.

6.3.2 Orbit analysis

We used orbit analysis to analyse the structure of researchers' neighborhood. Figure 6.4 shows the average orbit count for the NECTEC researchers network for graphlets up to size four, when compared to the Watts-Strogatz and Barabási models. In the first case, it can be noticed that orbits that are part of triangle-like graphlets (3, 10, 11, 12, 13, 14) and star-like graphlets (6, 7, 9, 10, 11) are more likely to be found in the researchers network, compared to a small world model. On the other hand, orbits that are part of chain-like graphlets (0, 1, 2, 4, 5, 8) are neither over or under represented. In the second case, orbits that are part of triangle-like graphlets are more likely to be found in the researchers network when compared to a preferential attachment model, while star-like graphlets and chain-like graphlets are under represented.

In a difference perspective, triangle-like graphlets are over represented in the researchers network, when compared to both null model. This is a natural consequence of the high level of transitivity in the network, and it shows that transitivity for this network is not merely a consequence of small world phenomenon or preferential attachment. This is also a consequence to the fact that researchers are clustered within the same projects. Star-like graphlets, instead, are highly represented in the network only when compared to the Watts-Strogatz models. This means that preferential attachment is sufficient to explain the presence of these types of graphlets. This may be due to the presence of high position researcher or project directors, that cover central positions in the network and collaborate with many other researchers on disjoint projects, that are therefore not connected to each other. Finally, star-like graphlets are under represented only when compared to the Barabási model. Since star-like configurations are a direct consequence of preferential attachment, their low presence in the research network shows that preferential attachment is not very strong. To summarize, triangle-like graphlets are more represented than star-like graphlets in the researcher network. This means that transitivity, more than preferential attachment, is the main force that drives researchers to connect. In other words, researchers that are choosing their collaborators prefer a common peer to a highly skilled person. Of particular interest, the graphlet G_6 (brokerage) is also highly represented in the network when compared to both models. Within this graphlet, orbit 11 serves as bridge between two connected nodes (orbit 10) and a single one (orbit 9). This orbit is specifically found in collaboration networks where actors that cover such position serve as "mediators" between different teams/groups.

We then proceeded by computing correlations between performance metrics and orbit counts. We used number of artifacts, number of projects and career length as productivity metrics and average IC score as quality metric (see Methods section). Figure 6.5 shows the correlation matrix for the NECTEC researchers network. The orbits that are positively correlated with productivity metrics are orbit 2, 5, 7, 8, 11, 13, 14. There is no specific graphlet type that stands out (triangle, star or chain). At the same time, degree is also not

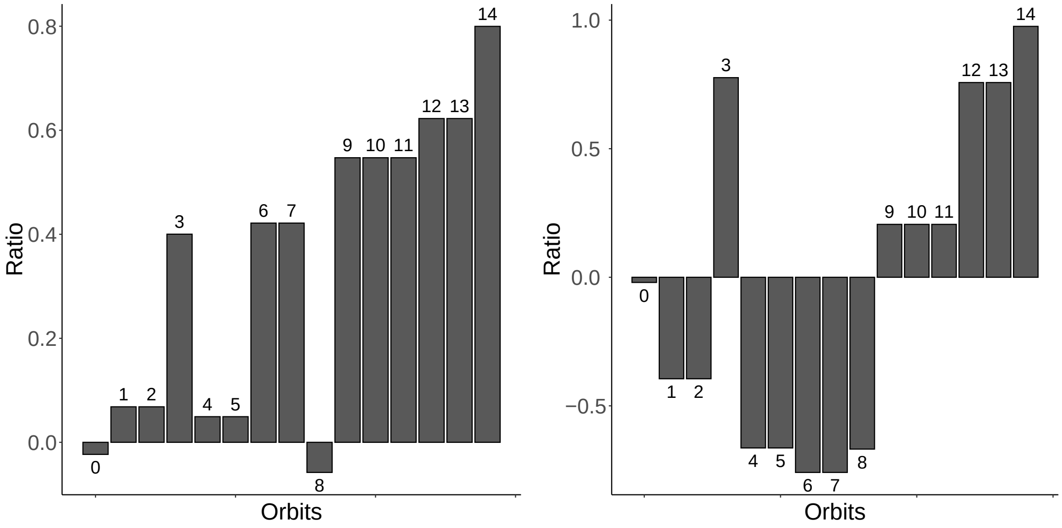


Figure 6.4: Orbit analysis, up to 4 nodes, for the researchers network. Bars go upwards if an orbit is more likely to be found in the real network compared to the null model. Left uses the Watts-Strogatz model as null model, while right uses Barabási model.

the key factor. In fact, among all orbits with degree two, some correlate positively (orbits 2, 5, 8, 12) and some others don't (orbits 3, 10). It can be noticed, instead, that all orbits that correlate positively with productivity metrics are the ones that, within the graphlet, are the most central. This result is in agreement with the previous correlation analysis, which showed that central nodes (nodes with high betweenness centrality) are more productive. For what concerns the IC score, instead, there is no correlation with any of the orbits analyzed.

6.3.3 Team evolution

We analyzed the distribution of team size and lifespan over time for the NECTEC researchers network. Figure 6.6 shows the results. The network is composed of a few large teams, while the majority of the teams are small (i.e. less than ten members). Team lifespan, instead, shows a bi-modal distribution, with most teams having either a short or long lifespan.

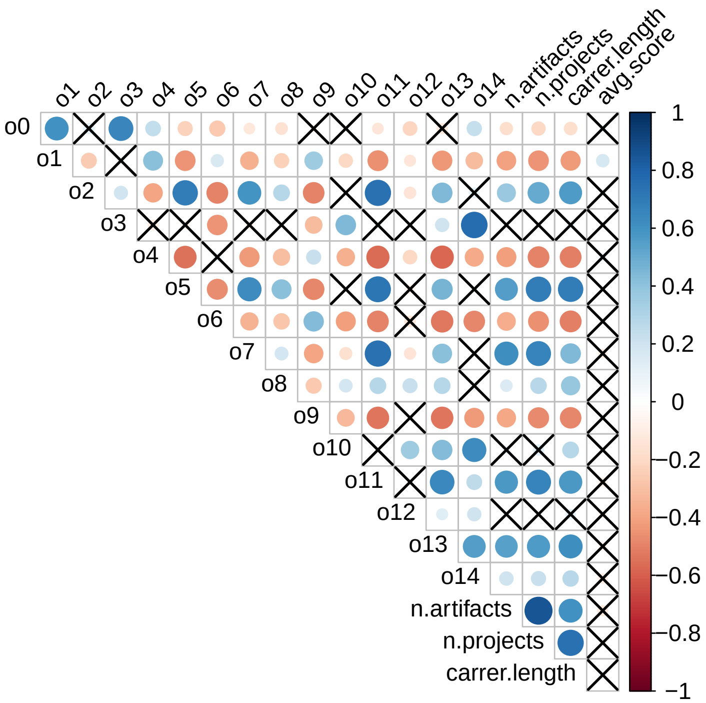


Figure 6.5: Correlation matrix for the NECTEC researchers network. Orbit number i is indicated as oi. Blue indicates positive correlation and red indicates negative correlation. Color intensity indicates the strength of correlation. Insignificant coefficients, according to p-value $p = 0.001$, are marked with a cross.

We looked at the evolution of teams over time. We tracked size (team structure metrics), as well as number of artifact and average IC score produced by all researchers (team performance metrics). For this analysis, we only considered teams whose lifespan was ten or higher, hence focusing on teams that fall within the second mode of the lifespan distribution. Figure 6.7 shows all the results. Generally, all teams start off as small and grow in size in the successive time slices, they then either keep their size or shrink towards the end of their life time. As it can be noticed, there are very different teams. For some teams (5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 19, 22), an increase/decrease in size is followed by an increase/decrease in number of artifacts and IC score. This can be the case

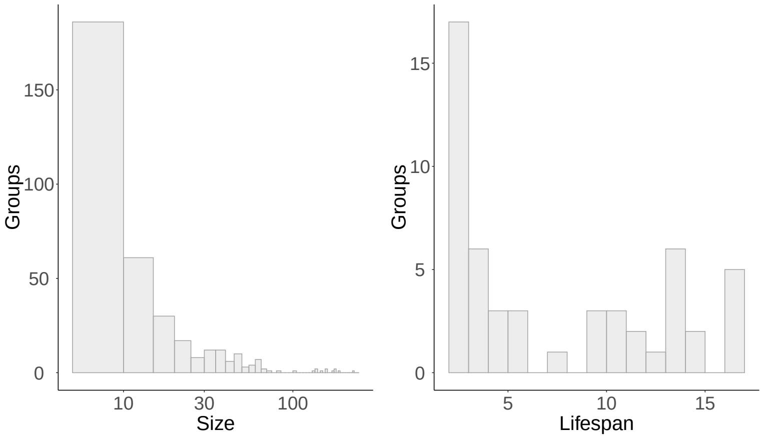


Figure 6.6: Team size and lifespan (measured in time slices) distribution for the temporal network. The diagram for team size has a logarithmic *x*-axes.

when experienced researchers join or leave a team, since their presence/absence highly affects the overall score of a team. In other cases (1, 2, 3, 4, 10, 16, 20, 21), an increase in size is followed by a decrease in number of artifacts and IC score. This can be the case when newcomers join an already established team, for which they do contribute in size but not to the overall score. Finally, an IC score that is significantly higher than the number of papers (2, 3, 10, 15) is an indicator of teams that produce, on average, high quality work.

For each team, we computed correlations between team performance metrics and team structure metrics in time. We used number of artifacts and average IC score as productivity and quality metrics. For team structure metrics, we used team size, autocorrelation, density, and ratio between outer and inner connections (see Methods section). For this analysis, we only considered teams whose lifespan was ten or higher. We then grouped together teams that showed similar results, forming three different groups. Groups have size of 12, 4 and 6 respectively. Figure 6.8 shows the correlation matrix for one representative team in each group. It can be noticed that team size is negatively correlated to performance metrics for group two and three, and positively correlated for group one. This indicates that a growth in terms of size does not always imply higher productivity or higher quality work. Autocorrelation has an influence only for the group three, correlating negativity with the performance metrics. This shows that, for these teams, high turnover is beneficial. Density correlates positively with number of artifacts and IC score for group two and three, while it has no effect for group one, showing that highly connected teams are more productive and produce higher quality work. Finally, the ratio between outer and inner connections is positively correlated to number of artifacts and IC score for all the three team groups. This means that teams that are not isolated, whose members engage in collaborations with members of other teams, are more productive and can produce higher quality work.

We also kept track and analysed all the changing events that affected the teams in the network such as teams growing or shrinking in size, the merging of small teams and splitting of large teams, and the birth and death of teams. Figure 6.9 shows all these events found in between time slices. Continuing and growing events are the most frequent in general, while merging, shrinking and splitting are more rare. The portion of forming and dissolving events corresponds to new projects starting off and terminating. Finally, Figure 6.10 gives a more visual perspective on the evolution of teams. It shows all teams of size 5 or larger, and how team members move from one team to another. It can be noticed that teams are generally growing over time. Small teams merge to form bigger ones and big teams maintain or grow their size over time.

6.4 Conclusion

In this study, we used a dataset from the National Electronics and Computer Technology Center (NECTEC) in Thailand, where researchers collaborate on different projects to produce a range of artifacts (intellectual properties, prototypes and scientific articles). We built a collaboration network where researchers are connected if they worked together on one or more artifacts.

First, we analyzed the distribution for some of the attributes of the NECTEC researchers networks, showing the skewed distribution of quantities related to performance

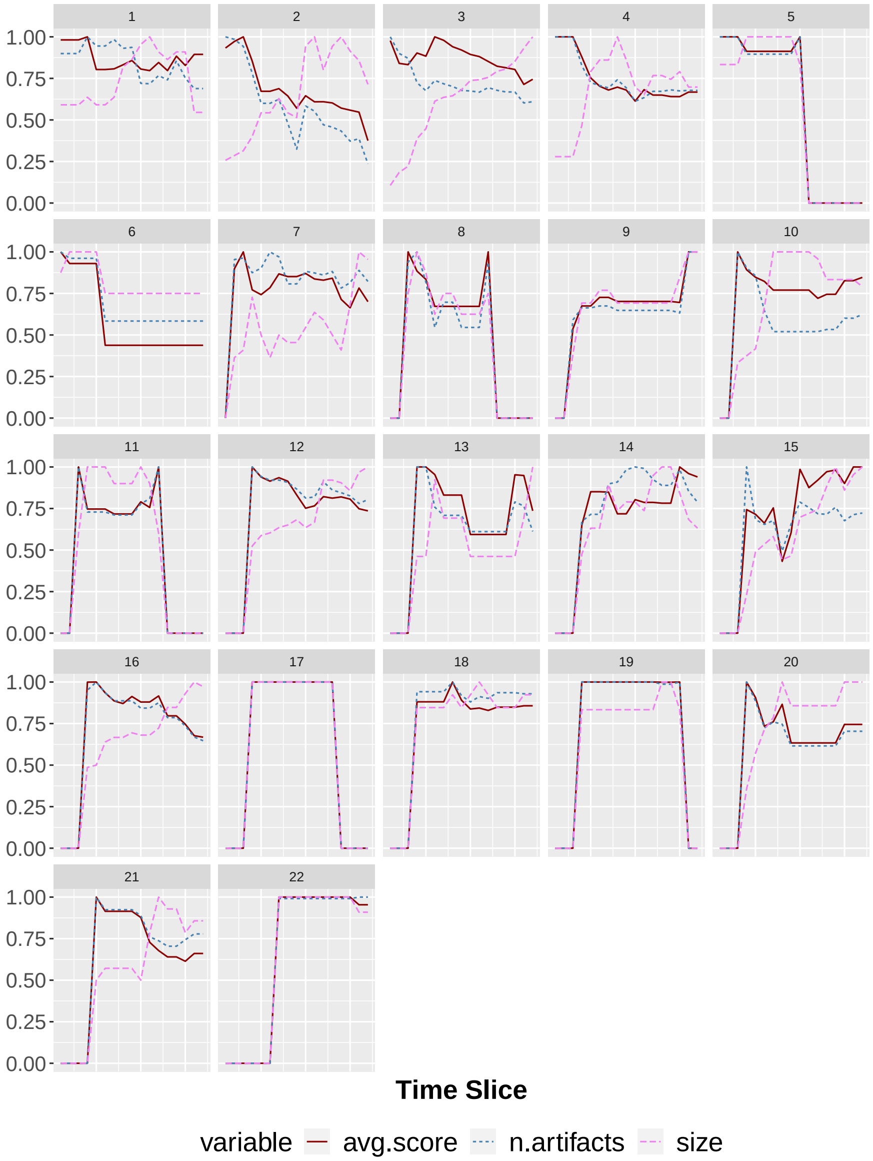


Figure 6.7: Each plot shows team size, number of artifacts and average IC score of a team over time. Only teams with lifespan ten or larger are considered. Time slice on the *x*-axes and observed variables (in percentage w.r.t maximum value) on the *y*-axes. Color and line type represents size, number of artifacts and average IC score.

such as number of artifacts produced, IC score and career length. We focused on measuring productivity and quality of research and development, while linking these metrics to the structure of the collaboration network. We have found that researchers that cover more central positions in the network, reflected by high betweenness centrality, are more productive. More productivity indicates e.g. high number of artifacts produced, engagement in multiple

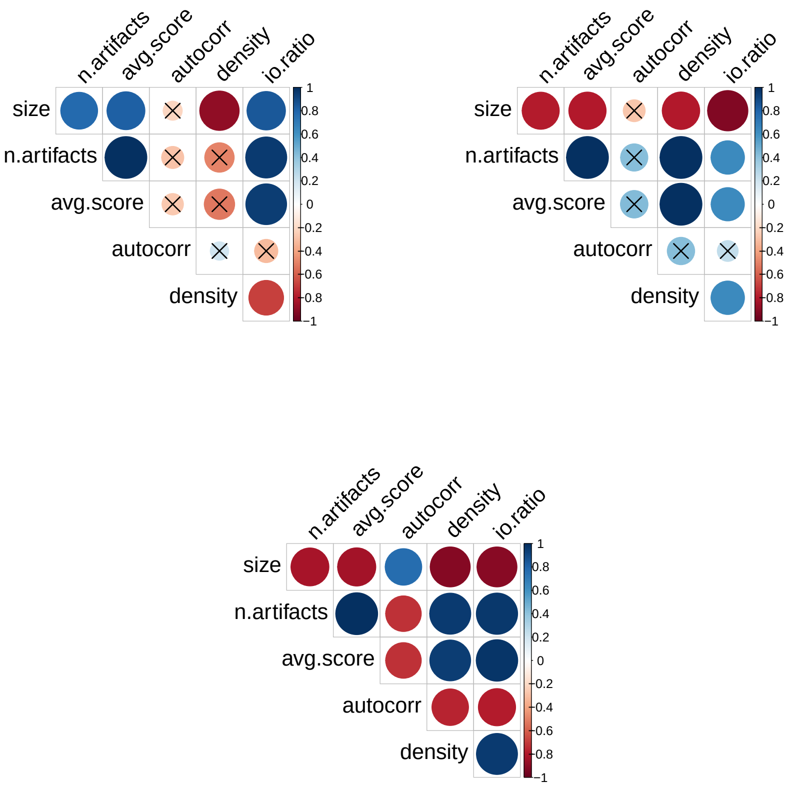


Figure 6.8: Correlation matrix for teams. One matrix for each group of teams is shown. Blue indicates positive correlation and red indicates negative correlation. Color intensity indicates the strength of correlation. Insignificant coefficients, according to p-value $p = 0.001$, are marked with a cross. Only teams whose lifespan is ten or higher are considered.

projects or longer career span. At the same time, centrality metrics are not found to be correlated with average IC score, which measure quality of work rather than quantity. On the other hand, we found that researchers who have more collaborators, reflected in higher degree centrality, are more productive, but their average IC score is lower, therefore producing lower quality work. These results are in agreement with the work of Uddin, Hossain and Rasmussen who linked betweenness and degree centrality to productivity metrics [120]. These results are also in partial agreement with the findings of Feng and Alec [117], who showed that researchers engaging in more collaborations have a longer career and higher citation count. Using orbit analysis, we showed how triangle-like graphlets are more rep-

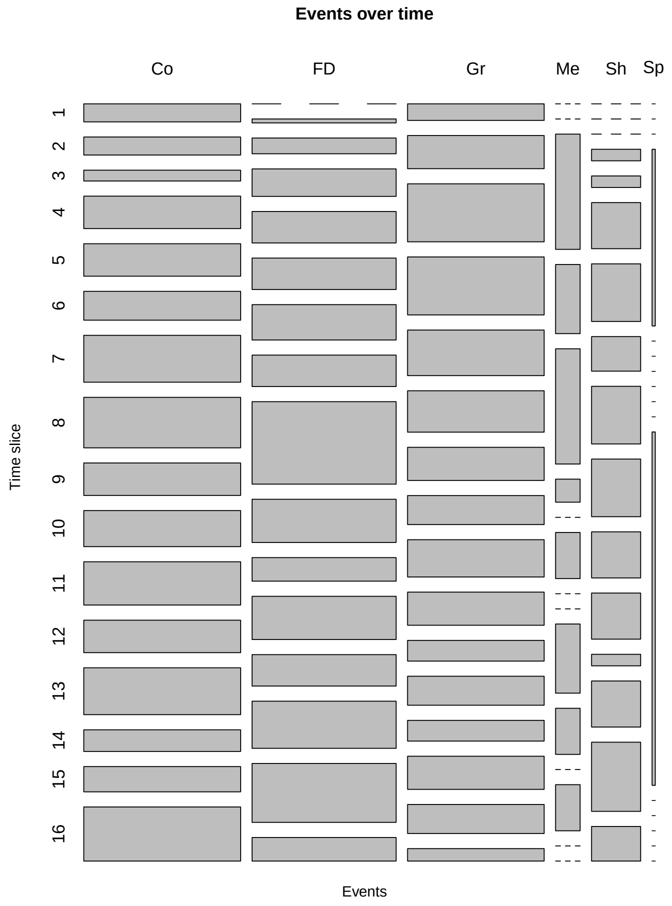


Figure 6.9: Events over time. Co: continuing, FD: forming or dissolving, Gr: growing, Me: merging, Sh: shrinking, Sp: splitting. For each column/event, the width of its rectangles corresponds to the frequency of this specific event over the whole time horizon. For each row/time slice, the height of a rectangle corresponds to the frequency of the event in the time slice when compered to the whole horizon.

resented than star-like graphlets in the researcher network, meaning that transitivity, more than preferential attachment, is the main force that drives researchers to connect. In other words, researchers that are choosing their collaborators prefer a common peer to a highly skilled person. Using correlation analysis, we also found that researchers who cover more central orbits withing a certain graphlet are more productive, result that is in agreement with the latter analysis.

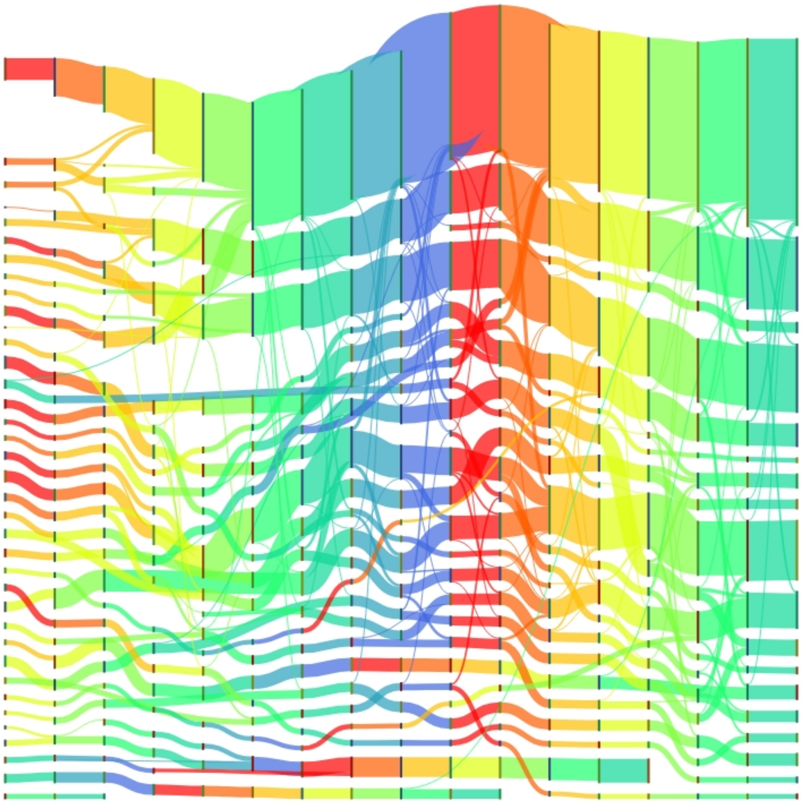


Figure 6.10: Evolution of teams over time for the NECTEC researchers network. Nodes represent communities at a certain time t, size represent team size. Edges represents researchers moving from one community at time t to another community at time $t+1$. Edge thickness represents the number of researchers. Colors represent different time transitions. Only communities of size 5 or greater are considered.

For what concerns teams, we analyze the distribution of team size and lifespan over time for the NECTEC researchers network, showing that the network is composed of a few large teams, while the majority of the teams are small (i.e. less than ten members). Team lifespan, instead, shows a bi-modal distribution, with most teams having either a short or long lifespan. We looked at the evolution of teams over time by tracking size (team structure metrics), as well as artifact and average IC score produced by all researchers (team performance metrics). We showed that the tracking of this quantities allows to identify teams that produce high quality work, as well as to show teams' turnover, particularly newcomers/experienced teams members that join/leave a team. We also computed correlation between team performance metrics (number of artifacts and IC score) and team structure metrics (team size, autocorrelation, density, ratio between outer and inner connections). We found that highly dense (connected) teams are more productive and produce higher quality work. This result is in agreement with Reagans and Zuckerman's work [129], who confirmed the closure view of social capital, stating that teams that experience more frequent communication among their members (higher density) can achieve higher productivity. High turnover is also beneficial and the result is in agreement with the work of Palla et al. [126], who showed how large teams last longer and perform better when turnover is high. Furthermore, teams that are not isolated, whose members engage in collaborations with members of other teams, are more productive and can produce higher quality work. For what concerns team size, a larger team does not necessarily imply higher productivity or quality. Finally, we tracked all changing events affecting teams in time, showing that teams are mostly growing in size, while other events like splitting an merging of teams are more rare.

6.4.1 Originality of the study and limitations

What makes this study original is the analysis based on a score that measures research quality rather than quantity (e.g. number of output and citation count), as well as a new set of structural metrics that help identify features of teams that are linked to success. Even though the analysis is limited to a rather small dataset, it is rich in information, especially for the availability of a quality score, which is often neglected or is just not available for analysis in other studies. Nonetheless, the methodology adopted can be extended to larger datasets/networks. The key findings of this study indicate that the success of a research institute needs to be assessed in the context of not just researcher or team level, but also on how the researchers engage in collaboration as well as on how teams evolve.

CHAPTER VII

SUMMARY

In this thesis, we discussed the importance of network analysis in the study of social complex systems. We showed how social network analysis can be used to address research questions in the humanities, particularly in history. Through a use case based on collaboration data in the field of Computer Science, we showed how social network analysis can be included in a research workflow, discussed the challenges that digital humanities is currently facing and the future opportunities for this growing fields. Furthermore, we showed how the emerging field of Science of Science uses social network analysis to analyze the factors the drive scientific success, both in academia and the private sector. For this purpose, we presented a use case based on collaboration data collected at the National Electronics and Computer Technology Center (NECTEC) in Thailand.

7.1 Contributions

This section will describe the contributions of this thesis. These can be summarized as follows:

1. We introduced the field of digital humanities, emphasizing the importance of creating a trading zone where scholars from different disciplines join forces and reflect on the epistemological and methodological challenges in Digital History. We discussed the

current challenges that the field is facing, such as the one of "drawing complicated graphs", "black boxes" and "data providers", as well as the ones related to data availability, storage and accessibility. We gave an overview of the most recent trends in historical network research. We emphasized the works that made best use of social network analysis in history, as well as the missed opportunities. We then presented the field of network analysis, providing a formalization of the concept of social networks, models that explain the mechanism governing complex networks and tools such as network metrics, orbit analysis and Exponential random graph models. In particular, we introduced the field of Science of Science and its motivation to discover, through network analysis, the hidden factors that drive scientific success.

2. We proposed MemLPA, a new version of the label propagation algorithm. It incorporates a memory element, in order for nodes to consider past states of the network in their decision rule. We gave an overview on community detection algorithms, focusing on the label propagation algorithm and its variations proposed in the literature. We investigated the advantages of memory and we found that its usage increases performance and prevents labels from overpropagating over the entire network. We conducted extensive experiments on the Lancichinetti–Fortunato–Radicchi benchmark, using normalized mutual information and adjusted rand index as performance metrics. We tested MemLPA against other existing label propagation algorithms that implement memory, showing that it provides better results. We also compared MemLPA to other well-known community detection algorithms to show that it outperforms some of them for values of the mixing parameter between 0.5 and 0.8. We conducted experiments on a set of real world networks of different nature, using modularity to evaluate the quality of the community assignments found, that further confirmed our finding. Finally, we performed a topological analysis using the Lancichinetti–Fortunato–Radicchi benchmark, comparing the topological properties of

the communities found to the ground-truth community structure. These results were achieved while keeping MemLPA completely scalable, using local interaction only and running in linear time.

3. We presented a use case, drawn from the collaboration with a historian colleague, showing how social network analysis can be used to answer research questions in history. We built a temporal co-authorship network based on a snapshot of the DBLP bibliographic database taken on 17 September 2015. The dataset contained gender information, based on first name. As for ethnicity, it was automatically retrieved based on family name and gender. We addressed the gender and ethnic bias problem in computer science research by looking at different collaboration patterns in the temporal co-authorship network. We started with the following research questions: "Do men and women, as well as researchers of different ethnic groups, show differences in collaboration patterns? How do these differences shape the network of collaboration?" We answered these questions by using network metrics that are based on researcher's position in the network and their neighborhood structure, such as betweenness, closeness, degree and clustering coefficient, rather than using classical performance metrics based on number of publications and number of citations. We computed these metrics for all researchers in the co-authorship network and ranked them from highest to lowest. We found that the women, being considered a minority in both size and in social perspective, score lower than men in terms of performance metrics and are more close-knit. Using orbit analysis, women are shown to occupy peripheral positions in the network, while men are more likely to cover central positions. Researchers of color, being considered a minority socially but larger in size when compared to white researchers (about 40% of them being Asian), score higher in ranks, while white researchers are more close-knit. Looking at the evolution of the co-authorship network, we showed that differences in gender and ethnicity are

narrowing over time. Furthermore, different historical periods had different trends. For example, women used to perform better in the middle of the 1980s and after the 2000s, while researchers of color are seeing an upward trend since the 1990s.

4. We presented a use case based on collaboration data collected at the National Electronics and Computer Technology Center (NECTEC) in Thailand, from 2009-10-03 to 2018-07-26. Researchers collaborate on different projects and team up to produce a range of artifacts (intellectual properties, prototypes and scientific articles). For each artifact, a score that measures quality of research is available and shared between the researchers that contributed to its creation, according to their percentage of contribution. We built a temporal collaboration network where researchers are connected if they worked together on one or more artifacts. We started with the following research questions: ”What are the collaborations patterns that lead to individual and group success?” We answered this question measuring productivity and quality of research and development, while linking these metrics to the structure of the collaboration network. We used number of artifacts, number of projects, number of collaborators and career length as performance metrics. As for network metrics, we used betweenness, closeness, and clustering coefficient. We found that researchers that cover more central positions in the network, reflected by high betweenness centrality, are more performing. At the same time, centrality metrics are not found to be correlated with average IC score, which measure quality of work rather than quantity. For what concerns teams, we used social groups dynamics to track their evolution over time. We proposed some team structural metrics that can be used to assess team’s performance. We used team size, autocorrelation, density and ratio between outer and inner connections as structural metrics. Using correlation analysis, we found that close-knit teams (high density), teams with high turnover (high

autocorrelation) and teams whose members actively seek for collaborations outside of their team (high outer/inner connection ratio) are the most successful teams.

7.2 Future challenges

Some potential areas for future efforts could include the following:

- MemLPA was developed and presented in Chapter IV to detect disjoint communities. The algorithm can be extended to detect overlapping groups, considering the nature of the memory elements to store information about multiple communities. Additionally, it would be beneficial to extend MemLPA to consider users' knowledge of the communities. In fact, especially in historical applications, prior knowledge of the communities is available and researchers are interested in taking this into account in the discovery of communities. Finally, the performance evaluation of MemLPA are assessed on static networks mainly. Additional experiments can be run to assess the performance of MemLPA in detecting communities in dynamical networks.
- All algorithms used to determine gender and ethnicity, including the ones used in Chapter V, are prone to bias. It would be beneficial, in order to have a more sound analysis, to use more algorithms and compare the different results to have a more clear idea on their limitations. Also a mixed analysis of gender and ethnicity can add a further layer of complexity and deeper understanding of the problem.
- The dataset provided by NECTEC, and analyzed in chapter VI, is rather small. Nonetheless, the centre is aware of the importance of the analysis and expressed the interest of enlarging the dataset and collecting more heterogeneous data, such as gender information of researchers. This can be a great opportunity to expand the analysis and use tool that otherwise would not be effective on such a small dataset.

REFERENCES

[1] T. van der Heijden, E. Andersen, J. Bronec, M. de Kramer, T. Durlacher, A. M. Fiscarelli, S. Haddadan, E. Kamlovskaya, J. Lotz, S. Mersch, *et al.*, "Presentation of the Luxembourg Centre for Contemporary and Digital History–C2DH Doctoral Training Unit," 2017.

[2] A. M. Fiscarelli, "Social network analysis for digital humanities: challenges and a use case," in *Digital History and Hermeneutics: Theory and Practice. Berlin* (J. Tatarinov and A. Fickers, eds.), De Gruyter, 2021.

[3] J. Tatarinov and A. Fickers, eds., *Digital History and Hermeneutics*. De Gruyter Oldenbourg.

[4] S. Wasserman, K. Faust, *et al.*, *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.

[5] B. H. Erickson, "Social networks and history: A review essay," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 30, no. 3, pp. 149–157, 1997.

[6] J. F. Padgett and C. K. Ansell, "Robust action and the rise of the Medici, 1400-1434," *American journal of sociology*, vol. 98, no. 6, pp. 1259–1319, 1993.

[7] N. Rosenthal, M. Fingrutd, M. Ethier, R. Karant, and D. McDonald, "Social movements and network analysis: A case study of nineteenth-century women's reform in New York State," *American Journal of Sociology*, vol. 90, no. 5, pp. 1022–1054, 1985.

[8] M. During, "The dynamics of helping behavior for Jewish refugees during the Second World War: The importance of brokerage," *Online Encyclopedia of Mass Violence*, 2016.

[9] T. Brughmans and U. Brandes, "Visibility network patterns and methods for studying visual relational phenomena in archeology," *Frontiers in Digital Humanities*, vol. 4, p. 17, 2017.

[10] J. E. Dobson, *Critical digital humanities: the search for a methodology*. University of Illinois Press, 2019.

[11] C. Lemercier, “Formal network methods in history: why and how?,” in *Social networks, political institutions, and rural societies*, pp. 281–310, 2015.

[12] J. E. Dobson, “Can an algorithm be disturbed?: Machine learning, intrinsic criticism, and the Digital Humanities,” *College Literature*, vol. 42, no. 4, pp. 543–564, 2015.

[13] R. Rosenzweig, “Scarcity or abundance? preserving the past in a digital era,” *The American historical review*, vol. 108, no. 3, pp. 735–762, 2003.

[14] C. Barats, V. Schafer, and A. Fickers, “Fading away... the challenge of sustainability in digital studies,” *Digital Humanities Quarterly*, vol. 14, no. 3, 2020.

[15] A. S. Breure and R. H. Heiberger, “Reconstructing science networks from the past,” *Journal of Historical Network Research*, vol. 3, pp. 92–117, 2019.

[16] P. O. Seglen, “The skewness of science,” *J. Am. Soc. Inform. Sci.*, vol. 43, no. 9, pp. 628–638, 1992.

[17] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, “Comparing community structure identification,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 09, p. P09008, 2005.

[18] Z. Yang, R. Algesheimer, and C. J. Tessone, “A comparative analysis of community detection algorithms on artificial networks,” *Scientific reports*, vol. 6, p. 30750, 2016.

[19] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[20] C. J. Anderson, S. Wasserman, and B. Crouch, “A p* primer: Logit models for social networks,” *Social networks*, vol. 21, no. 1, pp. 37–66, 1999.

[21] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, “An introduction to exponential random graph (p*) models for social networks,” *Social networks*, vol. 29, no. 2, pp. 173–191, 2007.

[22] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison, “Recent developments in exponential random graph (p*) models for social networks,” *Social networks*, vol. 29, no. 2, pp. 192–215, 2007.

[23] G. F. Riva, “Network analysis of medieval manuscript transmission,” *Journal of Historical Network Research*, vol. 3, pp. 30–49, 2019.

[24] M. Valleriani, F. Kräutli, M. Zamani, A. Tejedor, C. Sander, M. Vogl, S. Bertram, G. Funke, and H. Kantz, "The emergence of epistemic communities in the 'Sphaera' corpus: mechanisms of knowledge evolution," *Journal of Historical Network Research*, vol. 3, pp. 50–91, 2019.

[25] D. H. Cline, "Athens as a small world," *Journal of Historical Network Research*, vol. 4, pp. 36–56, 2020.

[26] A. J. Schauf and M. E. Varela, "Searching for hidden bridges in co-occurrence networks from Javanese wayang kulit," *Journal of Historical Network Research*, vol. 2, pp. 26–52, 2018.

[27] P. Galison, "Computer simulations and the trading zone," 1996.

[28] M. KEMMAN, *Trading zones of digital history*. PhD thesis, University of Luxembourg, 2019.

[29] S. VAN HERCK, *Visualising gender balance*. PhD thesis, University of Leuven, 2017.

[30] S. Van Herck and A. M. Fiscarelli, "Mind the gap. Gender and computer science conferences," in *IFIP International Conference on Human Choice and Computers*, pp. 232–249, Springer, 2018.

[31] S. van Herck, "Rebuilding the Office around the Mainframe: IBM's S/360 in context," 2019.

[32] S. van Herck, "Punched cards in acccounting at Helena Rubinstein," 2020.

[33] S. van Herck, "Gendered labour in business computing: the IBM S/360 in practice," 2019.

[34] S. Fiscarelli, Antonio Maria add van Herck, "Minorities in Computer Science. Gender and ethnic collaboration patterns in a temporal co-authorship network," *Plos one*, p. to appear in, 2021.

[35] R. Albert, H. Jeong, and A.-L. Barabási, "Internet: Diameter of the world-wide web," *nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[36] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[37] R. Albert, H. Jeong, and A.-L. Barabási, “Error and attack tolerance of complex networks,” *nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[38] N. Biggs, E. K. Lloyd, and R. J. Wilson, *Graph Theory, 1736-1936.* Oxford University Press, 1986.

[39] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks.* Cambridge university press, 2008.

[40] U. Brandes, “A faster algorithm for betweenness centrality,” *Journal of mathematical sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[41] A. Bavelas, “A mathematical model for group structures,” *Applied anthropology*, vol. 7, no. 3, pp. 16–30, 1948.

[42] A. Bavelas, “Communication patterns in task-oriented groups,” *The journal of the acoustical society of America*, vol. 22, no. 6, pp. 725–730, 1950.

[43] P. Erdős and A. Rényi, “On the evolution of random graphs,” *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[44] S. Milgram, “The small world problem,” *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[45] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’networks,” *nature*, vol. 393, no. 6684, p. 440, 1998.

[46] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[47] J. Scott, “Social network analysis,” *Sociology*, vol. 22, no. 1, pp. 109–127, 1988.

[48] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, “On modularity clustering,” *IEEE transactions on knowledge and data engineering*, vol. 20, no. 2, pp. 172–188, 2008.

[49] M. E. Newman, “Fast algorithm for detecting community structure in networks,” *Physical review E*, vol. 69, no. 6, 2004.

[50] R. Lambiotte, J.-C. Delvenne, and M. Barahona, “Random walks, markov processes and the multiscale modular organization of complex networks,” *IEEE Transactions on Network Science and Engineering*, vol. 1, no. 2, pp. 76–90, 2014.

[51] V.-L. Dao, C. Bothorel, and P. Lenca, “Estimating the similarity of community detection methods based on cluster size distribution,” in *International Conference on Complex Networks and their Applications*, pp. 183–194, Springer, 2018.

[52] G. K. Orman, V. Labatut, and H. Cherifi, “Comparative evaluation of community detection algorithms: a topological approach,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, no. 08, p. P08001, 2012.

[53] M. Jebabli, H. Cherifi, C. Cherifi, and A. Hamouda, “Community detection algorithm evaluation with ground-truth data,” *Physica A: Statistical Mechanics and its Applications*, vol. 492, pp. 651–706, 2018.

[54] S. M. Goodreau, J. A. Kitts, and M. Morris, “Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks,” *Demography*, vol. 46, no. 1, pp. 103–125, 2009.

[55] T. U. Grund and J. A. Densley, “Ethnic homophily and triad closure: Mapping internal gang structure using exponential random graph models,” *Journal of Contemporary Criminal Justice*, vol. 31, no. 3, pp. 354–370, 2015.

[56] T. Brughmans, S. Keay, and G. Earl, “Introducing exponential random graph models for visibility networks,” *Journal of Archaeological Science*, vol. 49, pp. 442–454, 2014.

[57] A. S. Breure and R. H. Heiberger, “Reconstructing science networks from the past,” *Journal of Historical Network Research*, vol. 3, no. 1, pp. 92–117, 2019.

[58] P. Holme and J. Saramäki, “Temporal networks,” *Physics reports*, vol. 519, no. 3, pp. 97–125, 2012.

[59] A. M. Fiscarelli, M. R. Brust, G. Danoy, and P. Bouvry, “A memory-based label propagation algorithm for community detection,” in *International Conference on Complex Networks and their Applications*, pp. 171–182, Springer, 2018.

[60] A. M. Fiscarelli, M. R. Brust, G. Danoy, and P. Bouvry, “Local memory boosts label propagation for community detection,” *Applied Network Science*, vol. 4, no. 1, pp. 1–17, 2019.

[61] M. E. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[62] A. Clauset, M. E. Newman, and C. Moore, “Finding community structure in very large networks,” *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[63] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 69, no. 10, p. P10008, 2008.

[64] P. Pons and M. Latapy, “Computing communities in large networks using random walks,” in *ISCIS*, vol. 3733, pp. 284–293, 2005.

[65] M. E. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Physical review E*, vol. 74, no. 3, p. 036104, 2006.

[66] J. Reichardt and S. Bornholdt, “Statistical mechanics of community detection,” *Physical Review E*, vol. 74, no. 1, p. 016110, 2006.

[67] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” *Physical review E*, vol. 76, no. 3, p. 036106, 2007.

[68] M. J. Barber and J. W. Clark, “Detecting network communities by propagating labels under constraints,” *Physical Review E*, vol. 80, no. 2, p. 026129, 2009.

[69] X. Liu and T. Murata, “Advanced modularity-specialized label propagation algorithm for detecting communities in networks,” *Physica A: Statistical Mechanics*, vol. 389, no. 7, pp. 1493–1500, 2010.

[70] I. X. Leung, P. Hui, P. Lio, and J. Crowcroft, “Towards real-time community detection in large networks,” *Physical Review E*, vol. 79, no. 6, p. 066107, 2009.

[71] J. Xie and B. K. Szymanski, “Community detection using a neighborhood strength driven label propagation algorithm,” in *2011 IEEE Network Science Workshop*, pp. 188–195, IEEE, 2011.

[72] L. Šubelj and M. Bajec, “Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction,” *Physical Review E*, vol. 83, no. 3, p. 036103, 2011.

[73] J. Xie and B. K. Szymanski, “Labelrank: A stabilized label propagation algorithm for community detection in networks,” in *Network Science Workshop (NSW)*, IEEE.

[74] S. Dongen, “A cluster algorithm for graphs,” 2000.

[75] J. Xie, B. K. Szymanski, and X. Liu, “Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process,” in *2011 ieee 11th international conference on data mining workshops*, pp. 344–349, IEEE, 2011.

[76] R. Hosseini and R. Azmi, “Memory-based label propagation algorithm for community detection in social networks,” in *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pp. 256–260, IEEE, 2015.

[77] F. Parés, D. G. Gasulla, A. Vilalta, J. Moreno, E. Ayguadé, J. Labarta, U. Cortés, and T. Suzumura, “Fluid communities: a competitive, scalable and diverse community detection algorithm,” in *International Conference on Complex Networks and their Applications*, pp. 229–240, Springer, 2017.

[78] J. Reginaldo Filho, M. R. Brust, and C. H. Ribeiro, “Consensus dynamics in a non-deterministic naming game with ahared memory,” *arXiv preprint arXiv:0912.4553*, 2009.

[79] T. G. Uzun, R. J. Da Silva-Filho, M. R. Brust, and C. HC, “Influence of shared memory and network topology in the consensus dynamics of a naming game,” in *XXXVIII Seminário Integrado de Software e Hardware (SEMISH). Anais do XXXI Congresso da Sociedade Brasileira de Computação*, 2011.

[80] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms,” *Physical review E*, vol. 78, no. 4, p. 046110, 2008.

[81] G. Csardi, T. Nepusz, *et al.*, “The igraph software package for complex network research,” *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[82] S. Van Herck and A. M. Fiscarelli, “Mind the gap: Gender and computer science conferences,” in *This Changes Everything – ICT and Climate Change: What Can We Do? 13th IFIP TC 9 International Conference on Human Choice and Computers, HCC13 2018, Held at the 24th IFIP World Computer Congress, WCC 2018, Poznan, Poland, September 19–21, 2018*, 2018.

[83] A. Swati, S. Ashish, M. Nitish, K. Rohan, and C. Denzil, “DBLP Records and Entries for Key Computer Science Conferences.” https://data.mendeley.com/datasets/3p9w84t5mr/, 2017. [Online; accessed January 11, 2018].

[84] T. J. Misa, *Gender codes: Why women are leaving computing*. John Wiley & Sons, 2011.

[85] J. Pizarro, "Race/ethnicity and justice in academia," *Race and Justice*, vol. 7, pp. 107–109, 04 2017.

[86] M. Sameer Khan, F. Lakha, M. Mei Jin Tan, S. Rajkumar Singh, R. Yu Chin Quek, E. Han, S. Mieng Tan, V. Haldane, M. Gea-Sánchez, and H. Legido-Quigley, "More talk than action: gender and ethnic diversity in leading public health universities," *Health Policy*, vol. 393, no. 10171, pp. 594–600, 2019.

[87] "Gendered innovations in Science, Health & Medicine, Engineering, and Environment," 2011-2018.

[88] G. Gutierrez y Muhs, Y. F. Niemann, C. G. Gonzalez, and A. P. Harris, eds., *Presumed Incompetent: The Intersections of Race and Class for Women in Academia*. Logan, UT: Utah State University Press, 2012.

[89] R. Padilla and R. Chávez, *The leaning ivory tower: Latino professors in American universities*. SUNY Series, United States Hispanic Studies, State University of New York Press, 1995.

[90] "Implicit bias in academia: a challenge to the meritocratic principle and to women's careers—and what to do about it," 2018.

[91] S. Agarwal, N. Mittal, R. Katyal, A. Sureka, and D. Correa, "Women in computer science research: What is the bibliography data telling us?," *ACM SIGCAS Computers and Society*, vol. 46, no. 1, pp. 7–19, 2016.

[92] S. Agarwal, N. Mittal, and A. Sureka, "Minority ethnic groups in computer science research: what is the bibliography data telling us?," *ACM SIGCAS Computers and Society*, vol. 47, no. 2, pp. 5–15, 2017.

[93] J. Huang, A. J. Gates, R. Sinatra, and A.-L. Barabasi, "Historical comparison of gender inequality in scientific careers across countries and disciplines," *arXiv preprint arXiv:1907.04103*, 2019.

[94] B. K. AlShebli, T. Rahwan, and W. L. Woon, "The preeminence of ethnic diversity in scientific collaboration," *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[95] W. Aspray, *Women and underrepresented minorities in computing: a historical and social study*. Springer, 2016.

[96] M. E. Newman, “Coauthorship networks and patterns of scientific collaboration,” *Proceedings of the national academy of sciences*, vol. 101, no. suppl 1, pp. 5200–5205, 2004.

[97] E. Elmacioglu and D. Lee, “On six degrees of separation in dblp-db and more,” *ACM SIGMOD Record*, vol. 34, no. 2, pp. 33–40, 2005.

[98] M. E. Newman, “Who is the best connected scientist? a study of scientific coauthorship networks,” in *Complex networks*, pp. 337–370, Springer, 2004.

[99] M. A. Nascimento, J. Sander, and J. Pound, “Analysis of SIGMOD’s co-authorship graph,” *ACM Sigmod record*, vol. 32, no. 3, pp. 8–10, 2003.

[100] D. Hiemstra, C. Hauff, F. de Jong, and W. Kraaij, “SIGIR’s 30th anniversary: an analysis of trends in IR research and the topology of its community,” in *ACM SIGIR Forum*, vol. 41, pp. 18–24, ACM New York, NY, USA, 2007.

[101] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, “Group formation in large social networks: membership, growth, and evolution,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 44–54, 2006.

[102] J. Huang, Z. Zhuang, J. Li, and C. L. Giles, “Collaboration over time: characterizing and modeling network evolution,” in *Proceedings of the 2008 international conference on web search and data mining*, pp. 107–116, 2008.

[103] C. Bird, P. Devanbu, E. Barr, V. Filkov, A. Nash, and Z. Su, “Structure and dynamics of research collaboration in computer science,” in *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 826–837, SIAM, 2009.

[104] M. Jadidi, F. Karimi, H. Lietz, and C. Wagner, “Gender disparities in science? dropout, productivity, collaborations and success of male and female computer scientists,” *Advances in Complex Systems*, vol. 21, no. 03n04, p. 1750011, 2018.

[105] G. Bravo-Hermsdorff, V. Felso, E. Ray, L. M. Gunderson, M. E. Helander, J. Maria, and Y. Niv, “Gender and collaboration patterns in a temporal scientific authorship network,” *Applied Network Science*, vol. 4, no. 1, p. 112, 2019.

[106] GenderizeAPI. https://genderize.io. [Online; accessed January 24, 2018].

[107] K. Imai and K. Khanna, “Improving ecological inference by predicting individual ethnicity from voter registration records,” *Political Analysis*, vol. 24, no. 2, pp. 263–272, 2016.

[108] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’networks,” *nature*, vol. 393, no. 6684, p. 440, 1998.

[109] N. Pržulj, D. G. Corneil, and I. Jurisica, “Modeling interactome: scale-free or geometric?,” *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[110] G. C. Bowker and S. L. Star, *Sorting things out: Classification and its consequences*. MIT press, 2000.

[111] A. M. Fiscarelli, M. R. Brust, R. Bouffanais, A. Piyatumrong, G. Danoy, and P. Bouvry, “Interplay between success and patterns of human collaboration: case study of a Thai research institute,” *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021.

[112] M. E. Newman, “Coauthorship networks and patterns of scientific collaboration,” *Proceedings of the national academy of sciences*, vol. 101, no. suppl 1, pp. 5200–5205, 2004.

[113] P. W. Mattessich and B. R. Monsey, *Collaboration: what makes it work. A review of research literature on factors influencing successful collaboration*. ERIC, 1992.

[114] G. J. Borjas and K. B. Doran, “Which peers matter? The relative impacts of collaborators, colleagues, and competitors,” *Rev. Econ. Stat.*, vol. 97, no. 5, pp. 1104–1117, 2015.

[115] A. M. Petersen, “Quantifying the impact of weak, strong, and super ties in scientific careers,” *P. Natl. Acad. Sci. USA*, vol. 112, no. 34, pp. E4671–E4680, 2015.

[116] G. Abramo, C. A. D’Angelo, and F. Di Costa, “The collaboration behavior of top scientists,” *Scientometrics*, vol. 118, no. 1, pp. 215–232, 2019.

[117] S. Feng and A. Kirkley, “Mixing patterns in interdisciplinary co-authorship networks at multiple scales,” *Sci Rep-UK*, vol. 10, no. 1, pp. 1–11, 2020.

[118] R. Cross, S. P. Borgatti, and A. Parker, “Making invisible work visible: Using social network analysis to support strategic collaboration,” *Calif. Manage. Rev.*, vol. 44, no. 2, pp. 25–46, 2002.

[119] A. M. Petersen, M. Riccaboni, H. E. Stanley, and F. Pammolli, "Persistence and uncertainty in the academic career," *P. Natl. Acad. Sci. USA*, vol. 109, no. 14, pp. 5213–5218, 2012.

[120] S. Uddin, L. Hossain, and K. Rasmussen, "Network effects on scientific collaborations," *PloS one*, vol. 8, no. 2, 2013.

[121] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Science*, vol. 3, no. 1, p. 9, 2014.

[122] J. Whitfield, "Group theory; What makes a successful team? John Whitfield looks at research that uses massive online databases and network analysis to come up with some rules of thumb for productive collaborations," *Nature*, vol. 455, no. 7214, pp. 720–724, 2008.

[123] S. Wuchty, B. F. Jones, and B. Uzzi, "The increasing dominance of teams in production of knowledge," *Science*, vol. 316, no. 5827, pp. 1036–1039, 2007.

[124] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, "Atypical combinations and scientific impact," *Science*, vol. 342, no. 6157, pp. 468–472, 2013.

[125] R. Guimera, B. Uzzi, J. Spiro, and L. A. N. Amaral, "Team assembly mechanisms determine collaboration network structure and team performance," *Science*, vol. 308, no. 5722, pp. 697–702, 2005.

[126] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, 2007.

[127] R. Kenna and B. Berche, "Critical masses for academic research groups and consequences for higher education research policy and management," *Higher Education Management and Policy*, vol. 23, no. 3, pp. 1–21, 2012.

[128] Y. Gao, Z. Zhu, R. Kali, and M. Riccaboni, "Community evolution in patent networks: technological change and network dynamics," *Applied network science*, vol. 3, no. 1, p. 26, 2018.

[129] R. Reagans and E. W. Zuckerman, "Networks, diversity, and productivity: the social capital of corporate r&d teams," *Organ. Sci*, vol. 12, no. 4, pp. 502–517, 2001.

[130] J. S. Coleman, "Social capital in the creation of human capital," *Am. J. Sociol.*, vol. 94, pp. S95–S120, 1988.

[131] R. S. Burt, *Structural holes: the social structure of competition*. Harvard university press, 2009.

[132] J. S. Coleman *et al.*, "Introduction to mathematical sociology," *Introduction to mathematical sociology.*, 1964.

[133] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3-5, pp. 75–174, 2010.

[134] P. Bródka, S. Saganowski, and P. Kazienko, "GED: the method for group evolution discovery in social networks," *Social Network Analysis and Mining*, vol. 3, no. 1, pp. 1–14, 2013.