Network science, big data analytics, and deep learning: An interdisciplinary approach to the study of citation, social and collaboration networks

Queen Mary
University of London

Yifan Qian

Submitted in partial fulfillment of the requirements

of the Degree of Doctor of Philosophy

School of Business and Management

Queen Mary University of London

28 September 2021

Statement of originality

I, Yifan Qian, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Yifan Qian

Date: 28 September 2021

Details of collaboration and publications:

I am the leading author of all the projects presented in the thesis. I conducted all the data collection, data analysis and prepared the manuscripts of related papers. My collaborators mainly played the role of supervision in these projects.

Chapters 6 and 7 have been published in leading peer-review journals in the fields of machine learning and data science (Qian et al., 2021a,b). The detailed information is shown below.

- **Yifan Qian**, Paul Expert, Tom Rieu, Pietro Panzarasa, and Mauricio Barahona. Quantifying the alignment of graph and features in deep learning. ***IEEE Transactions on Neural Networks and Learning Systems*** *(2021).* doi: https://doi.org/10.1109/TNNLS.2020.3043196

- **Yifan Qian**, Paul Expert, Pietro Panzarasa, and Mauricio Barahona. Geometric graphs from data to aid classification tasks with graph convolutional networks. ***Patterns Cell Press*** 2, no. 4 (2021): 100237. doi: https://doi.org/10.1016/j.patter.2021.100237

Abstract

Over the last few decades, networks have played an increasingly important role in multiple scientific domains, ranging from social science to physics and computer science. This thesis mainly focuses on three types of networks (citation networks, social networks, and collaboration networks) by combining theories and methods from network science, sociology, machine learning, and data science. Specifically, I present four projects concerned with two research clusters: social capital and deep learning. In the first project, I develop new measures of network effective size, i.e., intra- and inter-brokerage based on non-topological properties of nodes in directed and weighted networks, which can provide finer-grained perspectives on social capital. In the second project, I explore the social capital of cities extracted from the collaboration patterns of their resident scientists and their external collaborators by combining four large-scale bibliometric data sets. Results suggest that the relationship between the (internal or external) brokerage and scientific performance of cities is moderated by internal or external strong ties and the cities' geographical diversity. In the third project, I show that the classification performance of Graph Convolutional Networks (GCNs) is related to the alignment among features, graph, and ground truth, which I quantify using a subspace alignment measure corresponding to the Frobenius norm of the matrix of pairwise chordal distances between three subspaces associated with the three ingredients. The proposed measure is based on the principal angles between subspaces and has both spectral and geometrical interpretations. In the fourth project, I show that, if additional relational information is not available in the data set, one can improve classification by constructing geometric graphs from the features themselves and using them within a GCN. I also show that such feature-derived graphs increase the alignment of the data to the ground truth while improving class separation.

Acknowledgement

I would like first to express my sincere gratitude to my PhD supervisor, Prof. Pietro Panzarasa, who offered me a life-changing opportunity to become a computational social scientist and has been actively involved in all the projects that I worked on. I am deeply grateful for his unwavering support and belief in me at every stage of my PhD study.

I am also fortunate to collaborate with other scholars. I worked (i) with Prof. Noshir Contractor and Prof. Leslie DeChurch from Northwestern University on an empirical analysis of structural foundations of creativity in artistic industries using the novel brokerage measures I developed in Chapter 3; (ii) with Prof. Massimo Riccaboni from IMT Lucca and Dr. Luca Verginer from ETH Zürich on a project related to network foundations of the scientific performance of cities presented in Chapter 4; and (iii) with Prof. Mauricio Barahona and Dr. Paul Expert from Imperial College London on two projects concerned with Graph Neural Networks and classification tasks presented in Chapters 6 and 7. I wish to thank my mentors and collaborators for allowing me to learn interdisciplinary knowledge from them, ranging from network science to the science of science to machine learning with graphs.

Beyond these direct collaborations, I would like to thank Prof. Wenge Rong from Beihang University, Dr. Robert Peach, Dr. Zijing Liu, Dr. Sibo Cheng from Imperial College London, Mr. Xiancheng Li, Ms. Jianjian Gao, Ms. Ye Sun from Queen Mary, and Prof. Dashun Wang, Dr. Ching Jin and team members from the Centre for Science of Science and Innovation at Kellogg School of Management at

Northwestern University for their encouragement and insightful comments.

I also would like to thank Prof. Renaud Lambiotte from University of Oxford and Dr. Guven Demirel from Queen Mary for serving as my thesis examiners.

Last but not least, with great motion, I would like to express my special thanks to my wife, Ms. Xiaoqian Zheng, and my family for their continuous support for me. I cannot imagine where I would be without their love, guidance, patience, and care.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recently, as a result of the increasing availability of data and sophisticated big data analytical tools, a new area called "computational social science" (Lazer et al., 2009) has emerged lying at the intersection among computer science, network science, statistics, and the social sciences. On the one hand, social scientists are more interested in addressing substantive research questions associated with real-world applications. On the other, scientists from computational disciplines (e.g., computer science, physics, and mathematics) contribute new advanced data-driven computing and learning methods (e.g., artificial intelligence and network science), which can potentially be applied to real-world applications. The interplay between these two sources has been nurturing the development of computational social science.

Network science plays an important role in the study of various research topics within and beyond the computational social science. As economies and societies worldwide are becoming increasingly globalised, it is crucial to understand the systems where people, groups, and organisations interact with one another (Drucker, 1994; Harris, 2001). These systems can be represented as networks (called graphs in the mathematical literature), in which nodes (or vertices) are joined by links (or edges). The study of networks has attracted tremendous attention from various disciplines because networks are ubiquitous, and many research questions can be framed from a network-based perspective (Newman, 2018b; Vespignani, 2018; Wasserman and Faust, 1994; Watts, 2004). For instance, the Internet can be seen as a network in which the vertex is a computer or router, and the edge is a cable or wireless data connection. In a citation network, the vertex is an article, a patent, or a legal case, and the edge is a citation made by one vertex to another. In different disciplines, vertices and edges may be referred to in different ways. In sociology, vertices and edges are often called actors and ties, respectively. In computer science, they are often called nodes and links. As a highly interdisciplinary and fast-developing field, the study of networks is concerned with a variety of areas of investigation, including the network analysis of groups, institutions and social systems in sociology (Degenne and Forsé, 1999; Freeman, 2004; Lambiotte and Panzarasa, 2009), the analysis of complex systems in physics (Newman, 2003), graph theory in mathematics and computer science (Bollobás, 1998; Leskovec et al., 2005), machine learning with graphs in artificial intelligence (Bronstein et al., 2017), and network-based methods and theories in many other disciplines.

1.1 Part I: Social capital

Among the various scholars that have been interested in networks over the years, sociologists were among the first to propose network approaches to the study of social systems. One of the earliest works in which the concept of the network was proposed dates back to the 1930s (Roethlisberger and Dickson, 1939). Since then, sociologists have developed the social network perspective, which is characterised by an emphasis on the importance of social relationships among interacting actors (e.g., individuals, firms, and organisations) and by the attempt to systematically express theories, models, and applications in terms of relational concepts (Degenne and Forsé, 1999; Wasserman and Faust, 1994). A paradigmatic orientation shared by social network scholars is that social structure can be operationalised in terms of relations among actors and can be seen as emerging from the regularities or patterns generated by interactions among actors. In particular, a significant concern in the social sciences has been to understand or predict how individual behaviour is affected by the underlying social structure (Granovetter, 1985). To this end, sociologists have proposed and developed a number of theories and concepts, among which social capital (Coleman, 1988) plays a prominent role.

In the social sciences, social capital refers to the value that individuals or organisations can extract from their underlying social network structures within which they are socially embedded (Lin, 2002). It is widely acknowledged that social capital comes from the social network structure (Kilduff and Brass, 2010) which plays an essential role in maintaining or hindering a wide range of performance-based outcomes (Granovetter, 1977, 2005).

1.1.1 Intra- and inter-brokerage in social networks

Despite the general agreement on the salience of social structure for social capital, what kind of social network structure is more beneficial to actors is still controversial (Latora et al., 2013). In recent years, two opposing social structures based on the ego-centred networks of nodes have been proposed (see Figure 1.1): (i) closed structures in which the focal node (i.e., "ego") is mainly embedded in closed triangles; and (ii) open structures in which the ego is mainly surrounded by otherwise disconnected others. In this thesis, I shall focus on open structures through which a node is believed to have access to diverse views and information such that it can enjoy higher brokerage opportunities (Burt, 2009).

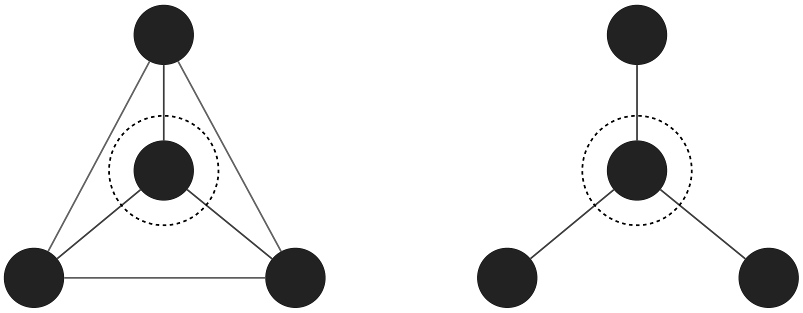


Figure 1.1: An illustration of closed (left) and open (right) structures. The focal node is surrounded by a dashed line.

The ideas related to open structures and brokerage have mainly been investigated by Ronald Stuart Burt, an American sociologist at the University of Chicago. To quantify the brokerage opportunities of a node within an open structure, Burt has proposed a set of network measures among which network effective size has been the one most widely used by scholars. Despite its popularity, it has been argued that effective size, by solely considering the network structure while ignoring the non-network attributes of the nodes, does not properly provide a comprehensive perspective on nodes' social capital (Aral and Van Alstyne, 2011; Fleming et al., 2007; Schilling and Fang, 2014; Shipilov and Li, 2008; Ter Wal et al., 2016; Uzzi, 1996). To address this limitation, in the project presented in Chapter 3, I propose two sets of new brokerage measures (intra- and inter-brokerage) as a function of a certain non-topological attributes of nodes by extending network effective size. The proposed measures can be applied to most general directed and weighted networks as well as to undirected and unweighted networks. This will allow us to quantify finer-grained measures of social capital based on combinations of topological and non-network attributes of nodes.

1.1.2 Network foundations of the scientific performance of cities

In recent years, the unit of analysis of studies related to social capital has been not only people or organisations but also geographical locations (Guan et al., 2015; Hristova et al., 2016). In particular, in the research communities of the "science of science" (Fortunato et al., 2018) and "research policy", several recent studies have explored the relationship between the social capital of countries or institutions extracted from scientific collaboration networks and their scientific performance (Cantner and Rake, 2014; Graf and Kalthaus, 2018; Guan et al., 2016). However, it is surprising to see that there are very few studies focusing on cities, considering the growing demand for more theoretical and empirical research of scientific collaboration at the city level (Neal, 2011). Also, previous related studies concerned with cities quantified the social capital of a city based on the inter-city scientific collaboration network (Guan et al., 2015). Here I argue that the inter-city collaboration network aggregated from the individual scientist level lacks the actual collaboration patterns of scientists within and across cities. To address these limitations, Chapter 4 presents an empirical study where I examine the relationship between sources of social capital (e.g., brokerage) of a city and its scientific performance using large-scale bibliometric data sets. I apply the geo-social network approach (Hristova et al., 2016) and measure the social capital of a city based on the collaboration patterns among resident scientists and external collaborators, respectively. The main results show that my proposed finer-grained measures of social capital (e.g., internal brokerage and external brokerage) can capture different perspectives of network collaboration patterns and may have distinct associations with scientific performance.

Measuring social capital and studying how it is associated with performance-based outcomes can be considered an important research topic in the network analysis of social systems, and more generally in computational social science. In this case, given a non-topological node attribute, it is assumed that its values are available for all nodes in the network. However, in many real-world data, for a non-topological node attribute, some nodes (even a large portion of nodes) may lack its values such that social capital measures using both network structure and non-topological node attributes (e.g., my proposed intra- and inter-brokerage measures) cannot be applied directly. Thus, for such a non-topological node attribute, it is important that these missing values can be predicted, which is a classification task in machine learning. While there are many paradigms (tools and techniques) of machine learning, deep learning plays one of the prominent roles at present, and indeed it will be the main focus of the second part of my thesis.

1.2 Part II: Deep learning

Deep learning can be viewed as part of a broader family of machine learning methods for discovering the representations and structures of the input data or signals needed for performing feature detection, classification tasks, or regression tasks. More specifically, deep learning is a collection of machine learning algorithms for modelling high-level abstractions in data through the use of model architectures, which are composed of multiple non-linear transformations (LeCun et al., 2015). Typically, deep learning has been successfully used to process various signals, such as speech, images, and videos, characterised by an underlying low-dimensional Euclidean structure.

The adoption of deep learning in the fields of network science and computational social science has been lagging behind until very recently, primarily because the non-Euclidean nature of the graph-structured data does not enable a straightforward definition of basic operations such as convolutions. Geometric deep learning (GDL), mainly realised by Graph Neural Networks (GNNs), refers to a fairly broad set of emerging techniques attempting to generalise deep neural models to graphs (Bronstein et al., 2017), thus extending notions from deep learning techniques to graph-structured data. Understanding and applying such advanced deep neural models interacted with graphs have been becoming very popular in both the machine learning community and other scientific communities since 2017, around the time I started my PhD. As many real-world problems and applications in the social sciences can be framed from a network perspective, GDL can provide promising techniques and tools to learn useful vector representations of nodes and facilitate the prediction tasks in the network analysis of social systems. GNN has witnessed success in a variety of research domains including computer vision (Landrieu and Simonovsky, 2018; Xu et al., 2017), natural language processing (Hamilton et al., 2017; Kipf and Welling, 2017; Peach et al., 2020; Veličković et al., 2018), traffic (Li et al., 2018b; Yu et al., 2018), recommendation systems (Monti et al., 2017; Ying et al., 2018), chemistry (Duvenaud et al., 2015; Gainza et al., 2020) and many other areas (Allamanis et al., 2018; Choi et al., 2017, 2018; Li et al., 2018c; Qiu et al., 2018; Zügner et al., 2018). For an in-depth review of GNNs, see Ref. (Wu et al., 2020).

1.2.1 Quantifying the alignment of graph and features

In this part of my thesis, I shall present two projects in the area of GDL. The work outlined in Chapter 6 is concerned with the conceptual understanding of Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017), a prominent GNN architecture[1]. GCN has been introduced to extend the notion of convolution to graph-structured data by leveraging and combining the structural information of a graph and the features of the nodes. GCN has been shown to perform extremely well on node classification tasks on well-known benchmarks in a semi-supervised learning setting. In this setting, GCN integrates three sources of information: the features of the nodes, the structure of relationships between the nodes (i.e., the graph), and the ground truth labels of nodes. GCN uses a subset (i.e., training set) of the labels, the full features, and the full graph to train a model, which is then used to predict the labels of the rest of the nodes (i.e., test set). A toy

[1] The GCN (Kipf and Welling, 2017) paper has been cited more than $9,000$ times until 16 July 2021.

example of using GCN to predict topics of papers in a citation network is shown in Figure 1.2. In this example, nodes are scientific papers. Each node is associated with a high-dimensional feature vector extracted from its semantic content. Nodes are connected if one paper cites another and the directions of the links are ignored. Each node is associated with a label representing its scientific topic.

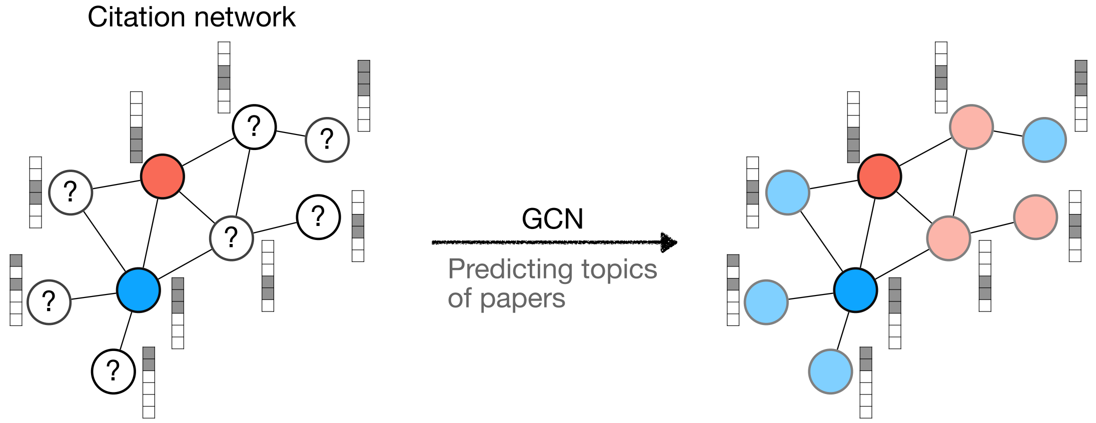


Figure 1.2: A toy example of using GCN to predict topics (colours of nodes) of papers (nodes) where papers are connected by citation links and each node is also associated with a feature vector representing its semantic content.

The implied assumption has been that the additional information provided by the graph will inevitably lead to improved classification accuracy compared with traditional graph-less classification methods. By contrast, my work shows and quantifies how, for GCN to perform successfully, there needs to be a degree of alignment among the three ingredients (features, graph, and ground truth) in the data set. In fact, in some cases, it might be beneficial to ignore the graph structure and use a simple graph-less approach such as the multilayer perceptron (MLP).

To characterise this phenomenon, I perform systematic randomisations of the graph structure and/or of the features that gradually erode the structure of the data set, and I show how to elucidate the relationship among, and the relative importance of, the three ingredients of GCN. My results thus shed light on what makes GCN perform better than simpler, limiting cases such as MLP, mean-field approaches, or ignoring the features. With this aim in mind, a key novelty of my study lies in the introduction of a subspace alignment measure, with spectral and geometric interpretations. I use the measure to assess the link between the alignment of the three ingredients and the classification performance of GCN and to quantify the alignment among graph, features, and ground truth that is needed for GCN to perform well. This also relates to the bias-variance trade-off in supervised machine learning algorithms. The higher alignment between ingredients corresponds to lower bias and higher variance, whereas the lower alignment between ingredients corresponds to higher bias and lower variance.

1.2.2 Geometric graphs from data to aid classification

The second project in Part II is presented in Chapter 7 where I show how GCN can be leveraged when dealing with data that have no explicit graph structure. Indeed, in a variety of empirical domains, the graph information is not readily available from the data sets. Here I show that, even in the typical case where only the sample features are available, it is still possible to extract a geometric graph from the feature vectors to encapsulate the closeness (i.e., the similarity) between samples and use this feature-derived graph within a GCN to improve the classification. Intuitively, the geometric graph acts as a conduit ensuring that class labels are predominantly shared between similar samples during learning.

To examine the improvement in classification performance induced by feature-derived graphs, I perform extensive computations on seven data sets from various disciplines using four well-known geometric graph construction methods. For each method and data set, I obtain geometric graphs of increasing edge density, and I show that there is a “sweet spot” in the density (neither over-sparse nor over-dense) where the GCN classification performance is maximised. My results show that classification aided by geometric graphs of appropriate density perform substantially better than classic graph-less classifiers (e.g., MLP) across my seven data sets. To gain further insight into the role played by the feature-derived geometric graphs, I quantify their effect using two measures. I show that geometric graphs with appropriate edge density both induce a subspace alignment of features and class membership vectors, as well as enhancing class separability.

Finally, I address the issue of how to make the graphs more efficient. Based on the well-known importance of spectral properties for graph partitioning (Lambiotte et al., 2014), I depart from purely geometric graphs and demonstrate that spectral sparsification of the selected feature-derived geometric graphs can improve further their classification performance while reducing the number of edges. Hence these efficient sparsified graphs combine a relatively sparse geometric graph that captures the local neighbourhood in turn sharpened by the global properties encapsulated in the Laplacian graph spectrum.

1.3 The structure of the thesis

As introduced above, my thesis can be articulated into two main building blocks: social capital and deep learning. I illustrate the structure of my thesis in Figure 1.3. In Part I, I will first introduce the background concerned with social capital in

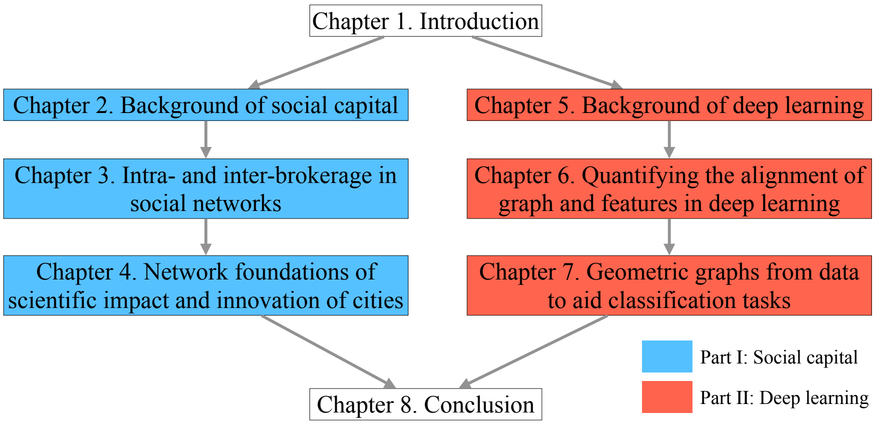

Chapter 1. Introduction
Chapter 2. Background of social capital
Chapter 5. Background of deep learning
Chapter 3. Intra- and inter-brokerage in social networks
Chapter 6. Quantifying the alignment of graph and features in deep learning
Chapter 4. Network foundations of scientific impact and innovation of cities
Chapter 7. Geometric graphs from data to aid classification tasks
Part I: Social capital
Part II: Deep learning
Chapter 8. Conclusion

Figure 1.3: The structure of the thesis.

Chapter 2, and then present two projects related to social capital in Chapters 3 and 4, respectively. In Part II, I will first introduce the background concerned with deep learning in Chapter 5, and then present two projects related to deep learning in Chapters 6 and 7, respectively. Finally, in Chapter 8, I will provide a summary of the main findings, discuss the contributions to the literature, and describe the agenda of my future research. Notice that notations in Part I and Part II are independent, and readers are cautioned not to confuse them.

Part I: Social capital

Chapter 2

Background

2.1 Structural foundations of social capital

The concept of social capital has long been a topic of debate across the social sciences. In general, the term “social capital” refers to the benefits that actors can derive from their relationships in families, communities, and other social networks. Although social capital has various definitions and there is currently no consensus on a specific one, it has been suggested that social capital might be seen as the “social glue” that brings people together and gives them a sense of belonging to this fast-growing and uncertain world (Catts and Ozga, 2005).

As Lin argued, the premise that seems to support most of the views of social capital is that investments in social relations can produce expected returns in the market, including the community, economic, financial, political, and labour markets (Lin, 2002). Social capital can have two different characteristics: it “inheres in the structure of relations between actors and among actors”, and “like other forms of capital, [it] is productive, making possible the achievement of certain ends that in its absence would not be possible" (Coleman, 1988).

Scholars agree on the salience of social structure for social capital, and in particular, they converge on the idea that individuals and organisations can gain information from their underlying networks (Kwon and Adler, 2014). However, what kind of social structure matters as a source of social capital is still a topic of debate and is controversy across the social sciences (Aral and Van Alstyne, 2011; Baum et al., 2012; Burt, 2005; Gargiulo and Benassi, 2000; Lin, 2002; Reagans and McEvily, 2003). In particular, over the past few years, two (apparently) opposite types of social structures have been proposed as possible sources of social capital: "closed" and "open" structures. Arguments in favour of both structures originate conceptually from Simmel's seminal theoretical contributions about the expansion of a dyadic relationship into a three-party relationship and the sociological significance of the third element (Simmel, 1950). Two functional roles have been suggested by Simmel which the third party can play in the triad: the mediator with the *tertius iungens* (or "the third who joins") orientation (Obstfeld, 2005) and the broker with the *tertius gaudens* (or "the third who enjoys") orientation (Burt, 2009). A toy example on the open and closed structures is shown in Figure 2.1.

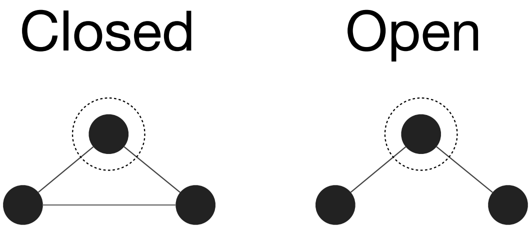


Figure 2.1: An illustration of closed (left) and open (right) structures in a triplet. The focal node is surrounded by a dashed line.

2.2 Closed structures

Advocates of the advantages of closed structures usually build their theory on Simmel's *tertius iungens* logic (Simmel, 1950) and Coleman's conception of social capital based on social cohesion (Coleman, 1988; Gamst, 1991). In particular, scholars have drawn upon the hypothesis that two separated actors sharing one common acquaintance are more likely to be connected than those without any common acquaintance (Davis, 1970; Davis et al., 1971; Holland and Leinhardt, 1971, 1977; Luce and Perry, 1949; Watts, 1999).

Generally, a closed structure refers to a densely connected network, rich in third-party relationships. It has been suggested that closed structures can induce trust (Burt and Knez, 1995; Gamst, 1991; Reagans and McEvily, 2003; Uzzi, 1997) and a sense of belonging (Coleman, 1988), maintain cooperative behaviour (Coleman, 1988; Ingram and Roberts, 2000) and social norms (Coleman, 1988; Gargiulo et al., 2009; Granovetter, 2005), and promote the creation of a common culture (Nahapiet and Ghoshal, 1998). Densely connected networks may also be related to clustering analysis, where the goal is to group samples such that samples in the same group are more similar (in some sense) to each other than to those in other groups (clusters). More specifically, given a network structure, the clustering analysis of nodes is commonly referred to as community detection.

Despite the benefits that closed structures can bring, actors in densely connected networks can also potentially bear a two-fold cost: local redundancy and social pressure. On the one hand, actors whose contacts are connected with each other are less likely to have access to diverse knowledge and resources than actors embedded in sparse networks (Granovetter, 1977). On the other hand, a cohesive structure can have a negative impact on actors because it can cause social pressure and induce actors to adopt similar beliefs and reach a unanimous consensus. As a result, it is likely that densely connected networks can promote the maintenance of the status quo instead of exploring new and diverging avenues (Fleming et al., 2007; Sosa, 2011).

2.3 Open structures

As both types of costs (i.e., local redundancy and social pressure) exist in cohesive structures, scholars have proposed an alternative conception of social capital associated with the benefits that actors can extract from participating in open structures. Generally, open structures refer to sparse networks rich in structural holes and brokerage opportunities (Burt, 2005, 2009, 2010; Lingo and O'Mahony, 2010; Stovel and Shaw, 2012). This conception is based on Simmel's description of the role of *tertius gaudens* (Simmel, 1950) in a triad, which is the role of the broker between otherwise disconnected others who intends to create and strengthen discontinuities in the social structure.

Open structures are believed to yield information advantages in the form of access to diverse views and information (Burt, 2004). Burt has thoroughly explored the idea that social capital can originate from brokerage opportunities related to open structures. The concept of "structural hole" is defined by Burt as a "separation between non-redundant contacts", "a relationship of non-redundancy between two contacts", "a buffer" that makes the two contacts "provide network benefits that are in some degree additive rather than overlapping" (Burt, 2009). Burt has suggested two sources of social capital related to structural holes: information benefits and control benefits. On the one hand, an actor embedded in an open structure rich in structural holes can combine different ideas and perspectives from neighbours who have weak connections with each other, and thus can come up with innovative ideas (Burt, 2004; Fleming et al., 2007; Sosa, 2011). On the other hand, control benefits are associated with the third party's ability to achieve an advantage by negotiating relationships with disconnected neighbours. An actor standing near a structural hole can control and transfer valuable information from one group to another, and ultimately combine various sources of information into new knowledge (Burt, 2009).

2.4 Empirical results

Since my work on social capital is concerned with open structures, here I shall review recent empirical results about applying the concept of brokerage in various scientific communities, including sociology, the economics of innovation, and research policy.

Ref. (Reagans and McEvily, 2003) examines the relationship between social cohesion and knowledge transfer using data from a contract R&D firm. Results suggest that higher brokerage of an individual is negatively associated with the willingness and motivation of individuals to invest time, energy, and effort in sharing knowledge with others. Ref. (Batjargal, 2007) studies the interaction effects of brokerage and experience of entrepreneurs on the performance of Internet ventures. The study is based on longitudinal surveys of 94 Internet ventures in Beijing, China. This study shows that the interaction of brokerage and the Western experience of entrepreneurs is positively related to the survival likelihood of Internet firms. In contrast, the interaction of brokerage and startup experience of entrepreneurs is negatively associated with firm performance. Ref. (Lu and McInerney, 2016) examines the cultural contingency of network structures in the contemporary Chinese academic labour market. Empirical results show that networks affording structural holes are only helpful for returnee's first promotion, whereas domestically trained PhDs benefit from network closure for obtaining their first promotion and subsequent promotions for all PhDs. Ref. (Guan and Liu, 2016) explores the association between brokerage and organisational innovations in terms of exploitation and exploration in the nano-energy field among 919 innovative organisations located in North America, Europe and Asia and 5107 observations during 2000–2013. This study argues that brokerage in a knowledge network hinders exploitative innovation but favours exploratory innovation. By contrast, brokerage in a collaboration network favours exploitative innovation but has a non-significant effect on exploratory innovation. Ref. (Guan et al., 2016) investigates the influence of collaboration network structure on national research and development efficiency on the country level. Results suggest that higher brokerage correlates positively with better future efficiency. Ref. (Liang and Liu, 2018) studies the evolution of government-sponsored collaboration and its impact on innovation in the Chinese solar photovoltaics sector. Results identify a positive relationship between brokerage and innovation performance, suggesting that an organisation should be embedded in open collaboration networks to increase its innovation performance. More recently, Ref. (Hur and Oh, 2021) studied the relationship between structural holes in the network of backward citations and future patent value using the United States pharmaceutical patent data. Specifically, this study shows that patents with less cohesive backward citation networks are likely to have higher private patent value and higher technological impact. In addition to extensive applications in sociology and economics, authors of Ref. (Hristova et al., 2016) have recently extended the concept of brokerage to the domain of geography. Specifically, they propose that a place that brokers between otherwise disconnected individuals in physical space can enjoy higher brokerage potential with respect to the social network of its residents and contacts outside the place. This concept can be naturally adapted to my context. In this case, the brokerage of a city can be expressed as its ability to connect otherwise disconnected scientists.

Chapter 3

Intra- and inter-brokerage in social networks

3.1 Introduction

To quantitatively measure and compare these two concepts of sources of social capital, i.e., closed and open structured introduces in Chapter 2, scholars have proposed two different measures: the local clustering coefficient (Watts, 1999; Watts and Strogatz, 1998) for closed structures, and the network effective size (Burt, 2009) for open structures. The local clustering coefficient quantifies the degree to which an ego node's alters tend to be connected with each other. In contrast, network effective size focuses on the absence of ties between alters of an ego node. If two alters are connected, it is believed that there is a certain portion of redundancy between them. Network effective size thus quantifies the non-redundant portion of the ego node's alters. The mathematical formalisations of local clustering coefficient and network effective size will be described in Section 3.2. It has been shown that these two measures have a simple mathematical relationship between them in undirected and unweighted networks (Latora et al., 2013). This relationship will be covered in Section 3.4. Although these two measures have been widely used in sociology to quantify closed and open structures, it has been argued that simply considering only the network structure while ignoring the non-network attributes of the actor does not provide a comprehensive perspective on the structural foundations of social capital (Aral and Van Alstyne, 2011; Fleming et al., 2007; Schilling and Fang, 2014; Shipilov and Li, 2008; Ter Wal et al., 2016; Uzzi, 1996).

Indeed so far little attention has been paid to proposing a new measure for quantifying open structures explicitly as a function of the non-topological attributes of the interacting nodes. To address this shortcoming, here I propose a formalisation of two new measures of brokerage – intra- and inter-brokerage – that can be seen as extensions of the formalisation of network effective size originally proposed by Burt for directed and weighted networks (Burt, 2004). Based on a certain non-topological attribute (described by a categorical variable, e.g., gender) A of nodes, I define the intra-brokerage of node i as the non-redundant portion of i's alters in i's ego-centred network that have the same attribute A as i. The inter-brokerage of node i, on the other hand, is here defined as the non-redundant portion of i's alters that are not characterised by the same attribute A as i. As a real-world example, I use a scientific collaboration network where nodes are scholars and two scholars are connected if they have co-authored at least one paper. If I consider gender as an attribute for each scholar, gender-based intra-brokerage of scholar

i will quantify non-redundant information one can receive from i's neighbours belonging to the same gender as i. In contrast, gender-based inter-brokerage can measure non-redundant information one can receive from i's neighbours with a different gender from i. Similarly, if the discipline is considered as a node attribute in a scientific collaboration network, discipline-based intra-brokerage of scholar i measures the brokerage opportunities one can obtain from other scholars within the same discipline. In contrast, discipline-based inter-brokerage quantifies the degree to which i can broker others from distinct disciplines. I will argue that defining such measures directly as a function of certain attributes of the actors will provide finer-grained perspectives on social capital.

The remainder of this chapter is organised as follows. In Section 3.2, I first review related concepts and formalisations regarding structural foundations of social capital. In Section 3.3, I introduce the node-level attribute-based brokerage measures, i.e., intra- and inter-brokerage, in directed and weighted networks. I then propose the simplified versions of intra- and inter-brokerage measures, and derive the relationship between these two measures and the intra- and inter-local clustering coefficient in undirected and unweighted networks in Section 3.4. In Section 3.5, I apply the new proposed intra- and inter-brokerage measures to a co-authorship network and compare them with standard brokerage measures introduced by Burt. Section 3.6 provides a discussion of my study. Finally, Section 3.7 summarises the contributions of my work to the literature.

3.2 Measuring social cohesion and structural holes

As social cohesion and structural holes represent two distinct sources of social capital, scholars have developed specific measures for properly detecting them: the local clustering coefficient and the effective size of a focal node's local (ego-centred) network, respectively. While the local clustering coefficient measures the extent to which a node is embedded in a closed cohesive structure, effective size uncovers the non-redundancy of a node's contacts, and thus can be used as an indicator of structural holes. In the rest of this section, I shall briefly review the formalisation of the local clustering coefficient and effective size.

Local clustering coefficient

Let me consider an unweighted and undirected network $G = (V, L)$ with a set of vertices V and a set of edges L, and let me focus on one of the nodes, node i. In order to measure the local cohesion of node i's ego-centred network, let me define $N(i)$ as the set of first neighbours of a node i and k_i as the number of nodes in $N(i)$. Formally, the local clustering coefficient of node i can be defined as (Watts, 1999; Watts and Strogatz, 1998):

$$C_i = \frac{l_i}{k_i(k_i - 1)/2} \tag{3.1}$$

where l_i represents the number of edges among i's neighbours. The local clustering coefficient of a node i is thus expressed as the ratio between the actual number of edges (i.e., l_i) and the maximum possible number of edges (i.e., $k_i(k_i - 1)/2$) between the nodes in i's ego-centred network. From another perspective, l_i can also be seen as the number of triangles in i's ego-centred network, and $k_i(k_i - 1)/2$ is equal to the maximum possible number of triangles centred on i. As a result, the local clustering coefficient represents the proportion of open triads centred on i that are closed into triangles, such that this measure is normalised between 0 and 1. On the one hand, C_i takes the minimum value of 0 in the case where there is no edge between any pair of i's neighbours. On the other hand, C_i takes the maximum value of 1 when all the neighbours of i are connected to each other, i.e., when the network formed by nodes in $N(i)$ is a complete graph.

Effective size

Effective size plays a key role among the various measures that Burt proposed to quantify the presence of structural holes and brokerage opportunities (Burt, 2009; Latora et al., 2013). The original formalisation of effective size was suggested by Burt based on the most general case of directed and weighted networks. Let me denote a directed and weighted network as $G = (V, L, W)$ where the corresponding weight associated with an edge (i, j) is represented by w_{ij}. The formula that Burt suggested for the effective size S_i of node i is defined as follows:

$$S_i = \sum_{j \in N(i)} \left[1 - \sum_{q \in N(i)} p_{iq} m_{jq} \right], \quad i \neq j \neq q \tag{3.2}$$

where j and q are two distinct nodes in $N(i)$. The term $\sum_{q \in N(i)} p_{iq} m_{jq}$ evaluates the extent to which j is redundant with respect to i's other neighbours. p_{iq} represents the proportion of i's network time and energy invested in the relationship with q, and m_{jq} is the marginal strength of j's relation with q. Formally, p_{iq} and m_{jq} are defined as follows:

$$p_{iq} = \frac{w_{iq} + w_{qi}}{\sum\limits_{t \in N(i)} (w_{it} + w_{ti})}, \quad i \neq q, i \neq t \tag{3.3a}$$

$$m_{jq} = \frac{w_{jq} + w_{qj}}{\max\limits_{r \in N(j)} (w_{jr} + w_{rj})}, \quad j \neq q, j \neq r \tag{3.3b}$$

where $\max_{r \in N(j)}(w_{jr} + w_{rj})$ is the largest of j's relations with any node in $N(j)$. Since $\sum_{q \in N(i)} p_{iq} m_{jq}$ represents the redundancy of j with respect to the other neighbours, $1 - \sum_{q \in N(i)} p_{iq} m_{jq}$ will refer to non-redundancy of node j. It can be noticed that the effective size is given by adding the non-redundancy of each node in i's ego-centred network. As a result, effective size is a measure which can detect the presence of structural holes and brokerage opportunities. The effective size of node i defined in Equation (3.2) varies from 1 to k_i, where k_i is the degree of node i. The ratio between effective size and k_i is called "efficiency", denoted as E_i, and ranges from 0 to 1 (and as such it can be seen as a normalised version of effective size).

3.3 Directed and weighted networks

3.3.1 Intra-brokerage

Let me consider a directed and weighted network $G(V, L, W)$ where the weight associated with an edge (i, j) is represented by w_{ij}. Table 3.1 reports the definitions of the variables that will serve as the ingredients for the intra-brokerage measure.

Table 3.1: Variables used to formalise the intra-brokerage measure

Symbol	Note
A	Attribute of a node
$N(i)$	The set of first neighbours of a node i
$N(i, A)$	A subset of $N(i)$ such that the attribute A of each node in $N(i, A)$ is the same as of node i
k_i^A	Number of elements in $N(i, A)$, i.e., $\|N(i, A)\|$
$D(i, \bar{A}, j)$	A subset of $N(i)$ such that: (i) each node in $D(i, \bar{A}, j)$ is connected with at least two nodes in $N(i, A)$ one of which is node j; and (ii) all nodes in $D(i, \bar{A}, j)$ do not share attribute A with node i
$D(i, \bar{A})$	$\bigcup_{j \in N(i,A)} D(i, \bar{A}, j)$
$d_i^{\bar{A}}$	Number of elements in $D(i, \bar{A})$, i.e., $\|D(i, \bar{A})\|$

Figure 3.1 illustrates an example of a simple network and the corresponding values for the variables in Table 3.2.

Table 3.2: Ingredients of intra-brokerage for the network in Figure 3.1

Variable	Example
A	Colour of a node
$N(1)$	$\{2, 3, 4, 5, 6, 7, 8, 9\}$
$N(1, A)$	$\{3, 4, 5, 6, 8\}$
k_1^A	5
$D(1, \bar{A}, j)$	$D(1, \bar{A}, 3) = \{\}$, $D(1, \bar{A}, 4) = \{9\}$, $D(1, \bar{A}, 5) = \{7, 9\}$, $D(1, \bar{A}, 6) = \{7\}$, $D(1, \bar{A}, 8) = \{\}$
$D(1, \bar{A})$	$\{7, 9\}$
$d_1^{\bar{A}}$	2

Based on the above variables and notations and on Burt's original formula for effective size, intra-brokerage S_i^{intra} can be defined as:

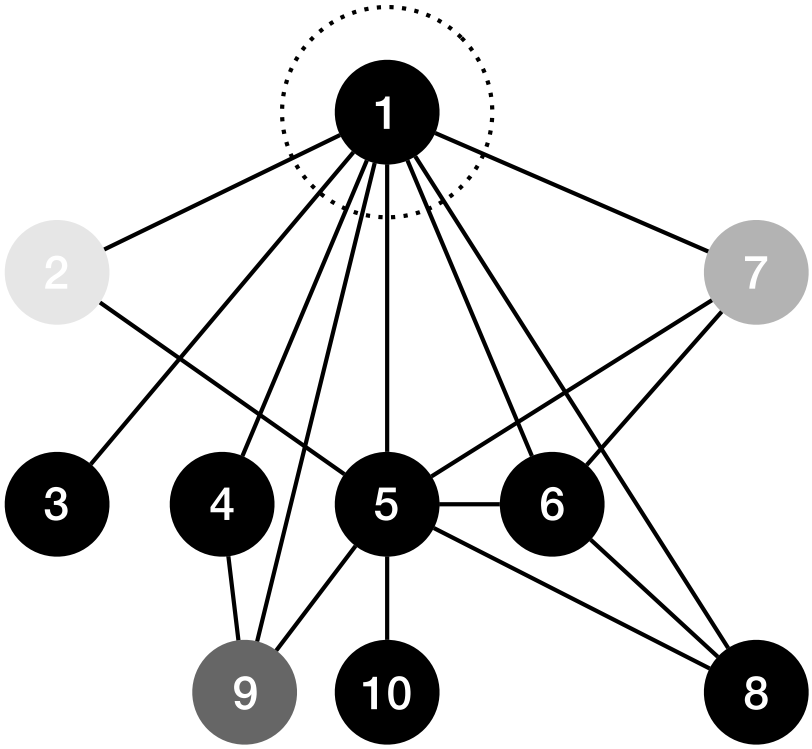


Figure 3.1: An example of intra-brokerage: Node 1 (ego) is circled by a dashed line, and the colour of each node represents the corresponding attribute. All edges are bidirectional in this example.

$$S_i^{intra} = \sum_{j \in N(i,A)} \left[1 - \sum_{q \in N(i,A) \cup D(i,\bar{A},j)} p_{iq}^{intra} m_{jq}^{intra} \right], \quad i \neq j \neq q \tag{3.4}$$

where p_{iq}^{intra} and m_{jq}^{intra} are defined as follows:

$$p_{iq}^{intra} = \frac{w_{iq} + w_{qi}}{\sum\limits_{t \in N(i,A) \cup D(i,\bar{A})} (w_{it} + w_{ti})}, \quad i \neq q, i \neq t \tag{3.5a}$$

$$m_{jq}^{intra} = \frac{w_{jq} + w_{qj}}{\max\limits_{r \in N(j)} (w_{jr} + w_{rj})}. \quad j \neq q, j \neq r \tag{3.5b}$$

The nodes in $D(i, \bar{A})$ are alters in i's ego-centred network that do not share attribute A with node i but connect with at least two nodes in $N(i, A)$. By including $D(i, \bar{A})$ in the redundancy part in Equation (3.4), the existence of such nodes in $D(i, \bar{A})$ will reduce the intra-brokerage opportunities of node i.

In Equation (3.4), $1 - \sum\limits_{q \in N(i,A) \cup D(i,\bar{A},j)} p_{iq}^{intra} m_{jq}^{intra}$ quantifies the non-redundant portion that i can receive from j. The final intra-brokerage of i is defined as the sum of the aforementioned non-redundant portion over all the neighbours j with the same class as i in terms of node attribute A.

Notice that the maximum value of S_i^{intra} is k_i^A. The ratio between S_i^{intra} and k_i^A is called "normalised intra-brokerage", denoted as E_i^{intra}, which ranges from 0 to 1.

3.3.2 Inter-brokerage

The formalisation of inter-brokerage mirrors closely the one for intra-brokerage. I will, therefore, show the proposed inter-brokerage measures without outlining the various steps in detail.

Table 3.3 reports the definitions of the variables that will be used to formalise inter-brokerage.

Table 3.3: Variables used to formalise the inter-brokerage measure

Symbol	Note
A	Attribute of a node
$N(i)$	The set of first neighbours of node i
$N(i, \bar{A})$	A subset of $N(i)$ such that each node in $N(i, \bar{A})$ does not share attribute A with node i
$k_i^{\bar{A}}$	Number of elements in $N(i, \bar{A})$, i.e., $\lvert N(i, \bar{A}) \rvert$
$D(i, A, j)$	A subset of $N(i)$ such that each node in $D(i, A, j)$: (i) is connected with at least two nodes in $N(i, \bar{A})$ one of which is node j; and (ii) all nodes in $D(i, A, j)$ share attribute A with node i
$D(i, A)$	$\bigcup_{j \in N(i,\bar{A})} D(i, A, j)$
d_i^A	Number of elements in $D(i, A)$, i.e., $\lvert D(i, A) \rvert$

Figure 3.2 illustrates an example of a simple network and the corresponding values for the variables in Table 3.4.

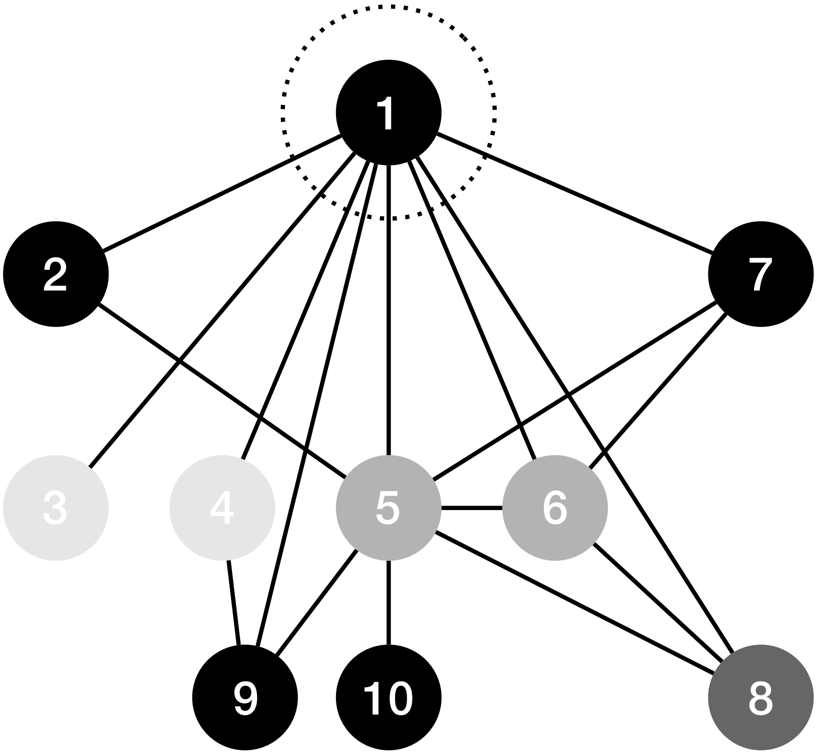


Figure 3.2: An example for inter-brokerage: Node 1 (ego) is circled by a dashed line, and the colour of each node represents the corresponding attribute. All edges are bidirectional in this example.

Table 3.4: Ingredients of inter-brokerage for the network in Figure 3.2

Variable	**Example**
A	Colour of a node
$N(1)$	$\{2, 3, 4, 5, 6, 7, 8, 9\}$
$N(1, \bar{A})$	$\{3, 4, 5, 6, 8\}$
$k_1^{\bar{A}}$	5
$D(1, A, j)$	$D(1, A, 3) = \{\}$, $D(1, A, 4) = \{9\}$, $D(1, A, 5) = \{7, 9\}$, $D(1, A, 6) = \{7\}$, $D(1, A, 8) = \{\}$
$D(1, A)$	$\{7, 9\}$
d_1^A	2

Inter-brokerage S_i^{inter} of node i within a directed and weighted network can now be defined as:

$$S_i^{inter} = \sum_{j \in N(i,\bar{A})} \left[1 - \sum_{q \in N(i,\bar{A}) \cup D(i,A,j)} p_{iq}^{inter} m_{jq}^{inter} \right], \quad i \neq j \neq q \tag{3.6}$$

where p_{iq}^{inter} and m_{jq}^{inter} are defined as:

$$p_{iq}^{inter} = \frac{w_{iq} + w_{qi}}{\sum\limits_{t \in N(i,\bar{A}) \cup D(i,A)} (w_{it} + w_{ti})}, \quad i \neq q, i \neq t \tag{3.7a}$$

$$m_{jq}^{inter} = \frac{w_{jq} + w_{qj}}{\max\limits_{r \in N(j)} (w_{jr} + w_{rj})}. \quad j \neq q, j \neq r \tag{3.7b}$$

The nodes in $D(i, A)$ are alters in i's ego-centred network that share attribute A with node i but connect with at least two nodes in $N(i, \bar{A})$. By including $D(i, A)$ in the redundancy part in Equation (3.6), the existence of such nodes in $D(i, A)$ will reduce the inter-brokerage opportunities of node i.

In Equation (3.6), $1 - \sum\limits_{q \in N(i,\bar{A}) \cup D(i,A,j)} p_{iq}^{inter} m_{jq}^{inter}$ quantifies the non-redundant portion that i can receive from j. The final inter-brokerage of i is defined as the sum of the aforementioned non-redundant portion over all the neighbours j belonging to a different class from i in terms of node attribute A.

Notice that the maximum value of S_i^{inter} is $k_i^{\bar{A}}$. The ratio between S_i^{inter} and $k_i^{\bar{A}}$ is called "normalised inter-brokerage", denoted as E_i^{inter}, which ranges from 0 to 1.

3.4 Simplified versions in undirected and unweighted networks

3.4.1 Intra-brokerage and the intra-local clustering coefficient

For an undirected and unweighted network, p_{iq}^{intra} and m_{jq}^{intra} can be simplified as:

$$p_{iq}^{intra} = \frac{1}{k_i^A + d_i^{\bar{A}}}, \quad i \neq q \tag{3.8a}$$

$$m_{jq}^{intra} = a_{jq}, \quad j \neq q \tag{3.8b}$$

where $a_{jq} = 1$ if node j is connected with node q, and $a_{jq} = 0$ otherwise. Therefore, intra-brokerage S_i^{intra} as defined in Equation (3.4) can be simplified as:

$$\begin{aligned} S_i^{intra} &= \sum_{j \in N(i,A)} \left[1 - \sum_{q \in N(i,A) \cup D(i,\bar{A},j)} p_{iq}^{intra} m_{jq}^{intra} \right] \\ &= \sum_{j \in N(i,A)} 1 - \sum_{j \in N(i,A)} \sum_{q \in N(i,A) \cup D(i,\bar{A},j)} p_{iq}^{intra} m_{jq}^{intra} \\ &= k_i^A - \sum_{j \in N(i,A)} \sum_{q \in N(i,A) \cup D(i,\bar{A},j)} \frac{a_{jq}}{k_i^A + d_i^{\bar{A}}} \\ &= k_i^A - \frac{1}{k_i^A + d_i^{\bar{A}}} \sum_{j \in N(i,A)} \sum_{q \in N(i,A) \cup D(i,\bar{A},j)} a_{jq}. \end{aligned} \tag{3.9}$$

In this case, where the network is undirected and unweighted, the intra-local clustering coefficient C_i^{intra} can be defined as:

$$C_i^{intra} = \frac{\sum\limits_{j \in N(i,A)} \sum\limits_{q \in N(i,A) \cup D(i,\bar{A},j)} a_{jq}}{k_i^A(k_i^A - 1) + k_i^A d_i^{\bar{A}}}, \quad \text{if} \quad k_i^A \geq 2. \tag{3.10}$$

where the numerator equals the sum of (i) two times of the number of edges between two nodes in $N(i, A)$; and (ii) the number of edges between two nodes where one is in $N(i, A)$ and the other is in $D(i, \bar{A})$. These are two types of redundancy considered in the definition of intra-brokerage in Equation (3.9). If intra-brokerage and intra-local clustering coefficient are considered as two related and opposing measures, the links contributing to the redundancy in intra-brokerage should become the numerator in the definition of intra-local clustering coefficient. The denominator corresponds to the maximum possible value associated with two types of edges in the numerator. In other words, intra-local clustering coefficient takes the maximum value when all nodes in $N(i, A)$ are connected and nodes between $N(i, A)$ and $D(i, \bar{A})$ are also all connected.

A simple relation between intra-brokerage and the intra-local clustering coefficient

In what follows, I develop the formal relationship between S_i^{intra} and C_i^{intra} for undirected and unweighted networks. Based on Equations (3.9) and (3.10), I obtain:

$$\begin{aligned}(k_i^A - S_i^{intra})(k_i^A + d_i^{\bar{A}}) &= k_i^A(k_i^A + d_i^{\bar{A}} - 1)C_i^{intra}\\ k_i^A - S_i^{intra} &= \frac{k_i^A(k_i^A + d_i^{\bar{A}} - 1)}{k_i^A + d_i^{\bar{A}}}C_i^{intra}\\ S_i^{intra} &= k_i^A - \frac{k_i^A(k_i^A + d_i^{\bar{A}} - 1)}{k_i^A + d_i^{\bar{A}}}C_i^{intra}\\ S_i^{intra} &= k_i^A - \frac{k_i^A(k_i^A + d_i^{\bar{A}} - 1)}{(k_i^A + d_i^{\bar{A}})(k_i^A - 1)}(k_i^A - 1)C_i^{intra}. \qquad (3.11)\end{aligned}$$

Equation (3.11) shows the relationship between intra-brokerage and the intra-local clustering coefficient for undirected and unweighted networks. This relationship is similar to Equation (3.12), which has been suggested for effective size and the local clustering coefficient in Ref. (Latora et al., 2013).

$$S_i = k_i - (k_i - 1)C_i. \qquad (3.12)$$

Notice that, in addition to Equation (3.12), Equation (3.11) contains one more coefficient, namely $\frac{k_i^A(k_i^A+d_i^{\bar{A}}-1)}{(k_i^A+d_i^{\bar{A}})(k_i^A-1)}$, which I will denote as α_i^A. Moreover, k_i^A in Equation (3.11) is equivalent to k_i in Equation (3.12). Using α_i^A and k_i^A, I can now simplify Equation (3.11) as:

$$S_i^{intra} = k_i^A - \alpha_i^A(k_i^A - 1)C_i^{intra}. \qquad (3.13)$$

In the case, where $d_i^{\bar{A}} = 0$, S_i^{intra} in Equation (3.13) can be further simplified as:

$$S_i^{intra} = k_i^A - (k_i^A - 1)C_i^{intra}. \qquad (3.14)$$

In the case where $d_i^{\bar{A}} \geq 1$ and $k_i^A \geq 2$, α_i^A is positive and larger than 1.

Bounds of intra-brokerage and the intra-local clustering coefficient

Based on the formalisations above, I will now examine the bounds of intra-brokerage and the intra-local clustering coefficient. According to Equation (3.9), the minimum value of S_i^{intra} corresponds to the case where each a_{jq} is equal to 1, and the maximum value of S_i^{intra} corresponds to the case where each a_{jq} is equal to 0:

$$\begin{aligned}
\min(S_i^{intra}) &= k_i^A - \frac{1}{k_i^A + d_i^{\bar{A}}} \sum_{j \in N(i,A)} \sum_{q \in N(i,A) \cup D(i,\bar{A},j)} 1 \\
&= k_i^A - \frac{k_i^A(k_i^A - 1 + d_i^{\bar{A}})}{k_i^A + d_i^{\bar{A}}} \\
&= \frac{k_i^A(k_i^A + d_i^{\bar{A}}) - k_i^A(k_i^A - 1 + d_i^{\bar{A}})}{k_i^A + d_i^{\bar{A}}} \\
&= \frac{k_i^A}{k_i^A + d_i^{\bar{A}}}, \text{ and}
\end{aligned}$$

$$\begin{aligned}
\max(S_i^{intra}) &= k_i^A - \frac{1}{k_i^A + d_i^{\bar{A}}} \sum_{j \in N(i,A)} \sum_{q \in N(i,A) \cup D(i,\bar{A},j)} 0 \\
&= k_i^A - 0 \\
&= k_i^A.
\end{aligned}$$

Thus,

$$\frac{k_i^A}{k_i^A + d_i^{\bar{A}}} \leq S_i^{intra} \leq k_i^A. \tag{3.15}$$

According to Equation (3.11), which shows the relationship between intra-brokerage and the intra-local clustering coefficient, I can obtain the bounds of the intra-local clustering coefficient as follows:

$$0 \leq C_i^{intra} \leq 1. \tag{3.16}$$

3.4.2 Inter-brokerage and the inter-local clustering coefficient

For an undirected and unweighted network, p_{iq}^{inter} and m_{jq}^{inter} can be simplified as:

$$p_{iq}^{inter} = \frac{1}{k_i^{\bar{A}} + d_i^{A}}, \quad i \neq q \tag{3.17a}$$

$$m_{jq}^{inter} = a_{jq}, \quad j \neq q \tag{3.17b}$$

where $a_{jq} = 1$ if node j is connected with node q, and $a_{jq} = 0$ otherwise. Therefore, inter-brokerage S_i^{inter}, as defined in Equation (3.6), can be simplified as:

$$S_i^{inter} = k_i^{\bar{A}} - \frac{1}{k_i^{\bar{A}} + d_i^{A}} \sum_{j \in N(i,\bar{A})} \sum_{q \in N(i,\bar{A}) \cup D(i,A,j)} a_{jq}. \tag{3.18}$$

In this case, where the network is undirected and unweighted, by mirroring C_i^{intra}, the inter-local clustering coefficient C_i^{inter} can be defined as:

$$C_i^{inter} = \frac{\sum\limits_{j \in N(i,\bar{A})} \sum\limits_{q \in N(i,\bar{A}) \cup D(i,A,j)} a_{jq}}{k_i^{\bar{A}}(k_i^{\bar{A}} - 1) + k_i^{\bar{A}} d_i^{A}}, \quad \text{if} \quad k_i^{\bar{A}} \geq 2. \tag{3.19}$$

A simple relation between inter-brokerage and the inter-local clustering coefficient

The relationship between S_i^{inter} and C_i^{inter} can be summarised as:

$$S_i^{inter} = k_i^{\bar{A}} - \frac{k_i^{\bar{A}}(k_i^{\bar{A}} + d_i^{A} - 1)}{(k_i^{\bar{A}} + d_i^{A})(k_i^{\bar{A}} - 1)}(k_i^{\bar{A}} - 1)C_i^{inter}. \tag{3.20}$$

Again, I can obtain a relationship that is similar to the one between effective size and the local clustering coefficient suggested in Ref. (Latora et al., 2013). Notice that, in addition to Equation (3.12), Equation (3.20) contains one more coefficient, namely $\frac{k_i^{\bar{A}}(k_i^{\bar{A}}+d_i^{A}-1)}{(k_i^{\bar{A}}+d_i^{A})(k_i^{\bar{A}}-1)}$, which I will denote as $\alpha_i^{\bar{A}}$. Moreover, $k_i^{\bar{A}}$ in Equation (3.20) is equivalent to k_i in Equation (3.12). Using $\alpha_i^{\bar{A}}$ and $k_i^{\bar{A}}$, I can now simplify Equation (3.20) as:

$$S_i^{inter} = k_i^{\bar{A}} - \alpha_i^{\bar{A}}(k_i^{\bar{A}} - 1)C_i^{inter}. \tag{3.21}$$

In the case where $d_i^{A} = 0$, S_i^{inter} in Equation (3.21) can be further simplified as:

$$S_i^{inter} = k_i^{\bar{A}} - (k_i^{\bar{A}} - 1)C_i^{inter}. \tag{3.22}$$

In the case where $d_i^{A} \geq 1$ and $k_i^{\bar{A}} \geq 2$, $\alpha_i^{\bar{A}}$ is positive and larger than 1.

Bounds of inter-brokerage and the inter-local clustering coefficient

Based on the formalisation of inter-brokerage and the inter-local clustering coefficient in undirected and unweighted networks, I will now examine the bounds of these two measures.

According to Equation (3.18), the minimum value of S_i^{inter} corresponds to the case where each a_{jq} is equal to 1, and the maximum value of S_i^{inter} corresponds to the case where each a_{jq} is equal to 0:

$$\min(S_i^{inter}) = \frac{k_i^{\bar{A}}}{k_i^{\bar{A}} + d_i^A}, \text{ and}$$

$$\max(S_i^{inter}) = k_i^{\bar{A}}.$$

Thus,

$$\frac{k_i^{\bar{A}}}{k_i^{\bar{A}} + d_i^A} \leq S_i^{inter} \leq k_i^{\bar{A}}. \tag{3.23}$$

According to Equation (3.20), which shows the relationship between inter-brokerage and the inter-local clustering coefficient, I can obtain the bounds of the inter-local clustering coefficient as follows:

$$0 \leq C_i^{inter} \leq 1. \tag{3.24}$$

3.5 A case study in a co-authorship network

3.5.1 Data set

I extracted the largest connected component from the co-authorship network in which the nodes are the scholars who published full papers in NeurIPS[1] 2016, and two nodes are connected if the corresponding scholars co-authored at least one accepted paper in NeurIPS 2016. I also assigned weights to edges based on the number of accepted papers two scholars co-authored in NeurIPS 2016 and the number of co-authors of each paper, as suggested in Ref. (Newman, 2001b). Specifically, let me begin by setting δ_i^p to denote whether scholar i is a co-author of paper p:

$$\delta_i^p = \begin{cases} 1 & \text{if scholar } i \text{ is a co-author of paper } p, \\ 0 & \text{otherwise.} \end{cases}$$

Then the weight w_{ij} of the edge between scholar i and scholar j is defined as:

$$w_{ij} = \sum_p \frac{\delta_i^p \delta_j^p}{n_p - 1}, \tag{3.25}$$

where n_p is the number of co-authors of paper p.

My final NeurIPS co-authorship network contains 75 nodes and 179 edges. I further collected information about the 75 scholars from their personal academic websites such that two attributes could be associated with each author. These two attributes are: (i) the scholar's gender; and (ii) the country where the scholar's affiliation is located. The reasons why gender and country are two important attributes

[1] NeurIPS (Conference and Workshop on Neural Information Processing Systems) is a flagship machine learning and computational neuroscience conference held every December.

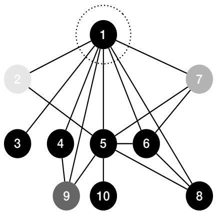
are as follows. On the one hand, recently, in the communities of the science of science and research policy, there is an increasing trend to study the demographic backgrounds of scholars in which gender is an essential one (Ni et al., 2021). Within- and across-gender collaboration might provide a scholar with different kinds of information. On the other hand, within- and across-country collaboration is a classic topic in the aforementioned two research communities (Wagner and Leydesdorff, 2005). A scholar might obtain different advantages from these two types of collaboration ties. Thus, it would be interesting to consider these two attributes and study the corresponding intra- and inter-brokerage.

Based on the above two node-level (non-topological) attributes, I can define two types of intra- and inter-brokerage measures: (i) gender-based; and (ii) country-based. Furthermore, depending on whether I consider the weights of the edges and whether I normalise the measures, I can evaluate the proposed intra- and inter-brokerage measures on this co-authorship network in four different cases: (i) the unweighted and unnormalised case; (ii) the weighted and unnormalised case; (iii) the unweighted and normalised case; and (iv) the weighted and normalised case. I also compare the proposed intra- and inter-brokerage measures with standard brokerage measures, i.e., effective size and efficiency. Section 3.5.2 summarises the empirical results.

3.5.2 Results

I summarise the notations used for the three sets of brokerage measures in the first three columns in Table 3.5. Each set contains four distinct measures: (i) unweighted and unnormalised; (ii) weighted and unnormalised; (iii) unweighted

Table 3.5: **Summary of notations used to formalise brokerage measures.** The weights of the edges of the schematic network are randomly assigned as follows: $(1, 2, 1)$, $(1, 3, 2)$, $(1, 4, 3)$, $(1, 5, 5)$, $(1, 6, 4)$, $(1, 7, 3)$, $(1, 8, 2)$, $(1, 9, 6)$, $(2, 5, 10)$, $(4, 9, 3)$, $(5, 6, 4)$, $(5, 7, 5)$, $(5, 8, 1)$, $(5, 9, 4)$, $(5, 10, 3)$, $(6, 7, 2)$, $(6, 8, 5)$, where the first two numbers refer to the two nodes connected by an edge and the last number is the weight of the edge. The network is considered to be undirected and weighted here.

Categories	Symbols	Notes	Toy network	Brokerage values of node 1
Standard brokerage (Burt, 2009)	S	Unweighted effective size		6.0000
	S^w	Weighted effective size		6.4103
	E	Unweighted efficiency		0.7500
	E^w	Weighted efficiency		0.8013
Intra-brokerage (This chapter)	S^{intra}	Unweighted and unnormalised intra-brokerage		3.5714
	$S^{w,intra}$	Weighted and unnormalised intra-brokerage		4.0440
	E^{intra}	Unweighted and normalised intra-brokerage		0.7143
	$E^{w,intra}$	Weighted and normalised intra-brokerage		0.8088
Inter-brokerage (This chapter)	S^{inter}	Unweighted and unnormalised inter-brokerage		2.2500
	$S^{w,inter}$	Weighted and unnormalised inter-brokerage		2.1111
	E^{inter}	Unweighted and normalised inter-brokerage		0.7500
	$E^{w,inter}$	Weighted and normalised inter-brokerage		0.7037

and normalised; and (iv) weighted and normalised. The first set is concerned with the more traditional measures of effective size and efficiency (Burt, 2009). The second and third sets of measures refer to intra- and inter-brokerage, respectively. In the last two columns of Table 3.5, I also report the brokerage values of node 1 in the schematic network in Figure 3.1.

I provide the code to compute my proposed measures of intra- and inter-brokerage at `https://github.com/haczqyf/brokerage`. The data set on the co-authorship network is also provided in the same Github repository.

Comparison of brokerage measures

First, I compute brokerage in terms of two distinct attributes (i.e., gender and country). As an example, Figure 3.3 shows a network where the focal attribute is gender, black nodes represent male scholars and grey nodes represent female scholars. As NeurIPS is a conference in the scientific field of computer science, scholars in this co-authorship network are mainly computer scientists. As shown in the figure, on the one hand, most scholars are male and only a minority are female. On the other, no two female scholars are directly connected. This results in a two-fold pattern: (i) in most ego-centred networks, alters tend to be male scholars; and (ii) female scholars' ego-centred networks tend to be characterised by larger inter-brokerage measures than intra-brokerage.

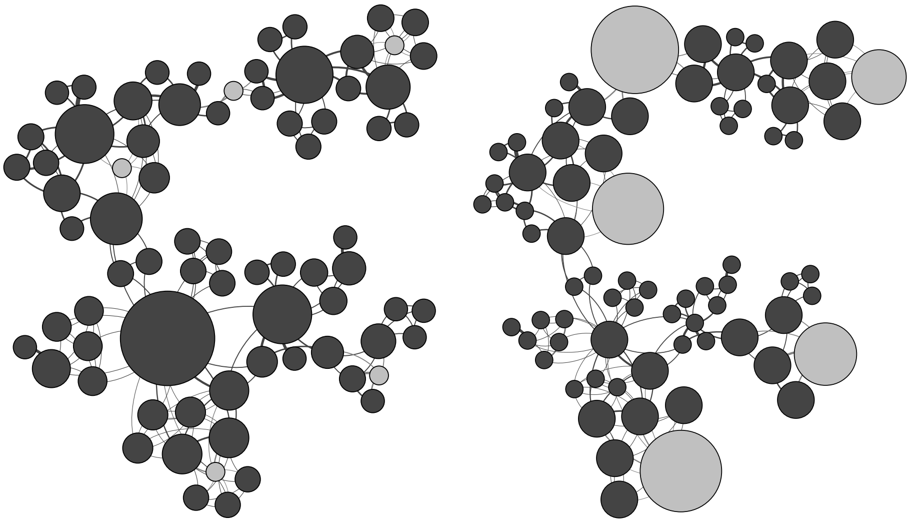
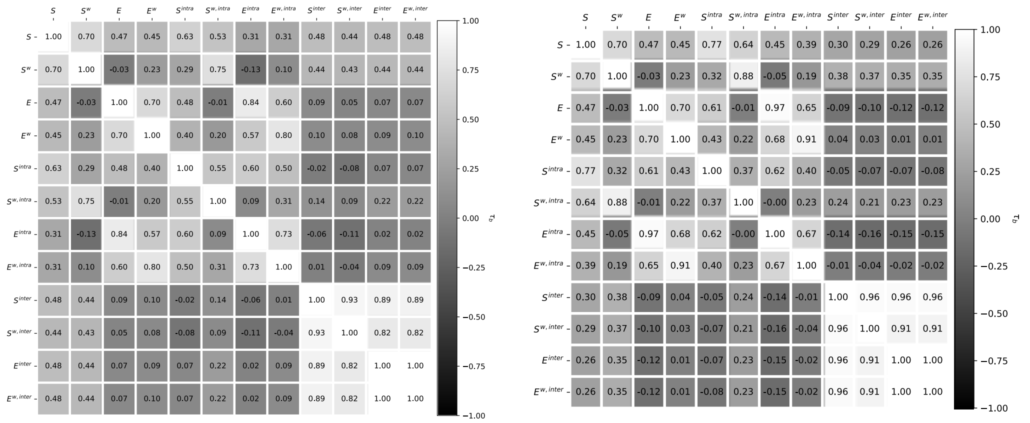

(a) Intra-brokerage ($S^{w,intra}$) (b) Inter-brokerage ($S^{w,inter}$)

Figure 3.3: Gender-based brokerage. Panels (a)-(b): Visualisation of the collaboration network. The colour of each node refers to its corresponding gender (black: male; grey: female), and the size of each node is proportional to the corresponding intra-brokerage (a) and inter-brokerage (b).

Second, for the three sets of brokerage, I calculate the Kendall's τ rank correlation coefficient (Kendall, 1938), denoted as $\tau_b(X, Y)$, which is a measure of the correspondence between two rankings X and Y. The Kendall's τ coefficient ranges from -1 to 1, such that values close to 1 indicate strong agreement and values close to -1 strong disagreement. Specifically, I compute the "τ-b" version of the Kendall's τ (Kendall, 1945) for each pair of rankings of nodes in the collaboration network introduced in Section 3.5.1 based on the brokerage measures summarised in Table 3.5. The Kendall's τ coefficients based on each of the two distinct attributes (i.e., gender and country) are visualised in Figure 3.4.



(a) Gender-based brokerage

(b) Country-based brokerage

Figure 3.4: The matrices of Kendall's τ coefficients of the brokerage measures. All correlations are statistically significant with p-values < 0.001.

I will now analyse the Kendall's τ rank correlation coefficients distinctively for each of the two different attributes. For each attribute, I can compare standard measures with each of the four versions of the proposed brokerage measures: unweighted and unnormalised, weighted and unnormalised, unweighted and normalised, and weighted and normalised. However, for the sake of clarity, I concentrate only on the weighted and unnormalised version, as the other three versions can be analysed in a similar way.

The first attribute I consider is gender, and the corresponding results are shown in Figure 3.4a. I found that $\tau_b(S^w, S^{w,intra}) = 0.75$, $\tau_b(S^w, S^{w,inter}) = 0.43$, and $\tau_b(S^{w,intra}, S^{w,inter}) = 0.09$. This indicates that there is relatively strong positive association for the pair $(S^w, S^{w,intra})$, and a weak positive association for the pair $(S^w, S^{w,inter})$, whereas there is a very weak positive association for the pair $(S^{w,intra}, S^{w,inter})$. This reflects the fact that the proposed intra- and inter-brokerage measures can potentially capture distinct information that would otherwise remain hidden if only the standard brokerage measures were applied.

In the real collaboration network, 5 scholars are female while 70 scholars are male.

Indeed, the distribution of classes of one attribute might influence the insights. To address this concern, I have now performed simulation analysis by increasing the number of females, and each scholar is randomly labelled as either male or female while keeping the network structure unchanged.

The number of females starts from 5. This number is increased by 5 every time until it reaches 35. When the number of females is small, gender is unbalanced. In contrast, when the number of females is 35, gender is almost well balanced. For each number of females considered, 100 times of random allocations are performed. Figure 3.5 shows the real network and two simulated networks with different numbers of females (5 and 35, respectively).

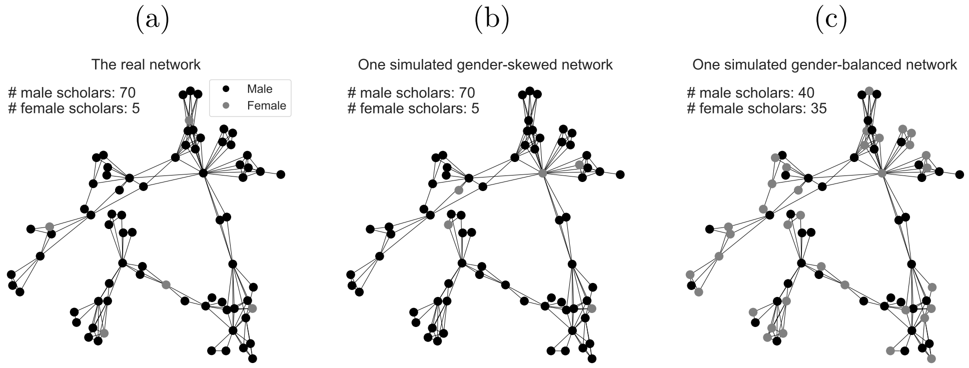
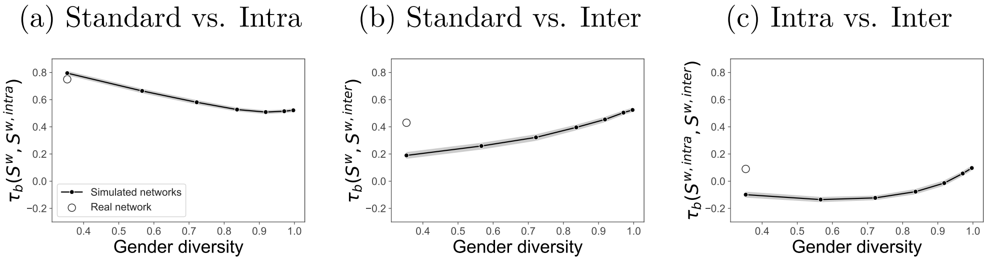


Figure 3.5: Visualisations of the real network and two simulated networks for the attribute gender.

For the real network and each simulated network, I also calculated two network-level measures: (i) *gender diversity*: it is calculated as the Shannon entropy (with base 2) of the gender distribution of scholars; and (ii) *gender heterophily*: it is calculated as the proportion of links for which two scholars have different genders. I also calculated the weighted unnormalised standard brokerage (S^w) and gender-based intra- and inter-brokerage ($S^{w,intra}$ and $S^{w,inter}$). The pairwise Kendall's τ correlation coefficients among them are also computed, denoted as

$\tau_b(S^w, S^{w,intra})$, $\tau_b(S^w, S^{w,intra})$ and $\tau_b(S^{w,intra}, S^{w,inter})$.

Figure 3.6 shows the relationships between these correlation coefficients and gender diversity. Results suggest that gender diversity is negatively correlated with $\tau_b(S^w, S^{w,intra})$, whereas it is positively correlated with $\tau_b(S^w, S^{w,intra})$ and $\tau_b(S^{w,intra}, S^{w,inter})$. Notice that the correlations among brokerage measures only vary within a small range when gender diversity changes. In particular, the correlation between intra- and inter-brokerage is always low, suggesting that my insights about intra- and inter-brokerage still hold even when gender is balanced in the network.



(a) Standard vs. Intra

(b) Standard vs. Inter

(c) Intra vs. Inter

Figure 3.6: The relationships between correlations among brokerage measures and gender diversity. The black dots show the means of correlation coefficients between two brokerage measures in simulated networks. The grey area represents the 95% confidence interval. The bigger grey empty circle corresponds to the real network.

I also plotted the relationship between correlation coefficients of brokerage measures and gender heterophily in Figure 3.7. Results show that gender heterophily is negatively correlated with $\tau_b(S^w, S^{w,intra})$, whereas it is positively correlated with $\tau_b(S^w, S^{w,intra})$ and $\tau_b(S^{w,intra}, S^{w,inter})$. Again, the correlation between intra- and inter-brokerage is low and only varies within a small range when gender heterophily changes, suggesting that my insights about intra- and inter-brokerage still hold even when gender heterophily is higher in the network.

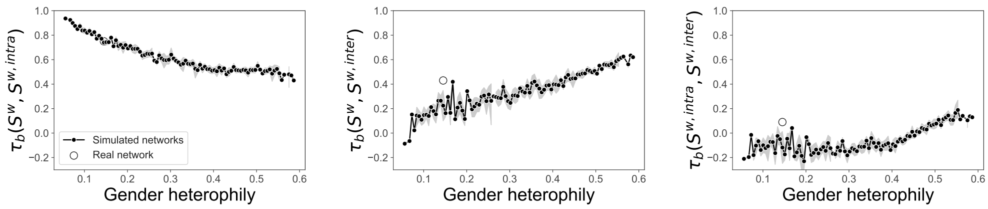

(a) Standard vs. Intra (b) Standard vs. Inter (c) Intra vs. Inter

Figure 3.7: The relationships between correlations among brokerage measures and gender heterophily. The black dots show the means of correlation coefficients between two brokerage measures in simulated networks. The grey area represents the 95% confidence interval. The bigger grey empty circle corresponds to the real network.

The second attribute of my study is country, and the corresponding results are shown in Figure 3.4b. Here, the country refers to the geographic location of the institution with which a scholar is affiliated. There are in total three different countries in the NeurIPS co-authorship network: the United States, Switzerland, and Canada. I found that $\tau_b(S^w, S^{w,intra}) = 0.88$, $\tau_b(S^w, S^{w,inter}) = 0.37$, and $\tau_b(S^{w,intra}, S^{w,inter}) = 0.21$, which again suggests non-overlapping rankings of nodes based on the different measures.

Among 75 scholars, 47 scholars are from the US, 17 scholars are from Switzerland, and 11 scholars are from Canada. If one country dominates the network, the standard brokerage and intra-brokerage tend to highly correlated. This also corresponds to the findings when the attribute is gender. The real collaboration network in terms of country is country-unbalanced. I first randomly allocated each scholar into one of the three countries by keeping the network structure and country distribution unchanged. 100 simulated networks were performed. Next, to consider the case where country distribution is balanced across three countries, I consider the case where the number of scholars from three countries is the same in the network, i.e., 25 from each country. Then in each simulated network, each scholar is assigned to one of the three countries. 100 simulated networks were performed. Like with gender diversity and gender heterophily, country diversity and country heterophily (two measures at the network level) are calculated for the simulated networks. In Figure 3.8, the real network and two simulated networks are shown in which one is country-unbalanced and the other is country-balanced.

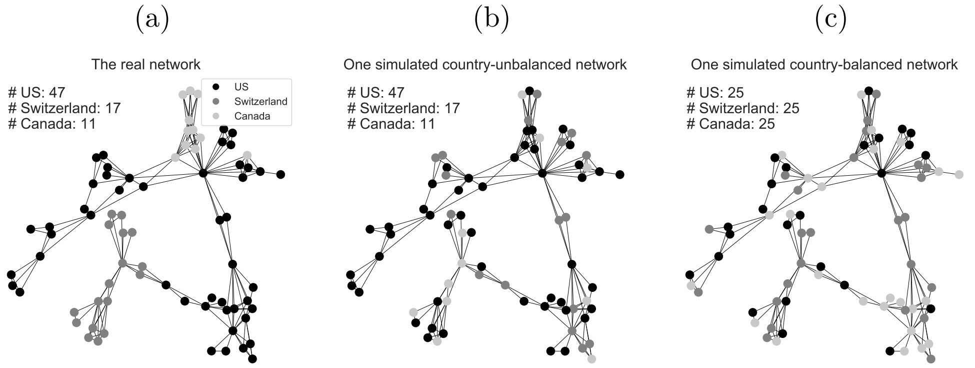


Figure 3.8: Visualisations of the real network and two simulated networks for the attribute country.

In Figure 3.9, I show the comparison between simulated country-unbalanced and country-balanced in terms of country diversity and country heterophily. As expected, country-balanced simulated networks have higher country diversity and country heterophily.

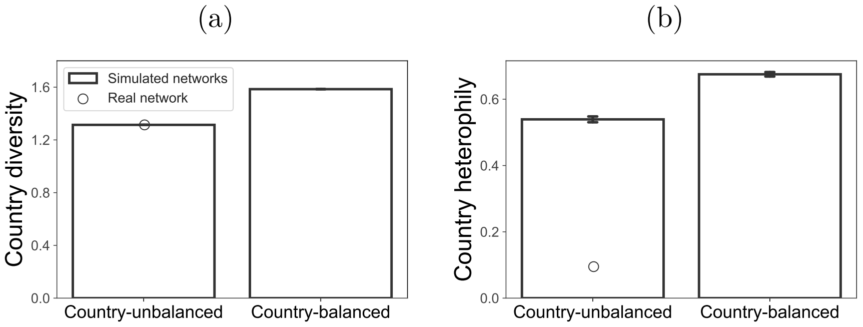
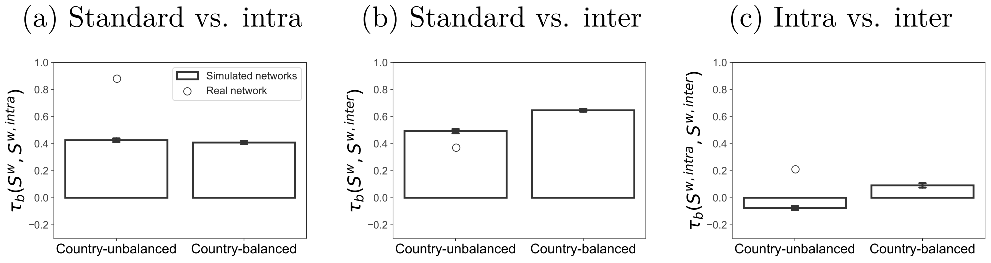


Figure 3.9: Bar plots show the means of The country diversity and country heterophily in simulated country-unbalanced and country-balanced networks. The error bar represents the 95% confidence interval. The bigger grey empty circle corresponds to the real network.

Like with gender, for each simulated network, I calculated the weighted unnormalised standard brokerage (S^w) and country-based intra- and inter-brokerage ($S^{w,intra}$ and $S^{w,inter}$). The pairwise Kendall's τ correlation coefficients among them are also computed. The mean of correlation coefficients of brokerage measures in country-unbalanced and country-balanced simulated networks are compared in Figure 3.10. Results suggest that, compared with country-unbalanced networks, in country-balanced networks, $\tau_b(S^w, S^{w,intra})$ tends to be lower whereas $\tau_b(S^w, S^{w,inter})$ and $\tau_b(S^{w,intra}, S^{w,inter})$ tend to be higher. In particular, even in country-balanced networks, the correlation between intra- and inter-brokerage is still low, which suggests that my insights about intra- and inter-brokerage can capture distinct perspectives of social capital still hold.


(a) Standard vs. intra

(b) Standard vs. inter

(c) Intra vs. inter

Figure 3.10: The correlations among brokerage measures in country-unbalanced and country-balanced networks. Bar plots show the means of correlation coefficients between two brokerage measures in simulated networks. The error bar represents the 95% confidence interval. The bigger grey empty circle corresponds to the real network.

The social capital of scholars: Brokerage and scientific impact

I further examine the association between my proposed measures and scholars' scientific impact in the co-authorship network. The scholar's scientific impact is measured by the number of citations that the scholar received between 2016 and 2018. I use OLS models with robust standard errors to regress the log of the scholar's citations on the standard and proposed intra- and inter-brokerage measures, respectively. Here I use weighted and unnormalised versions to exemplify my analysis. Moreover, covariates in regression models are standardised (with mean 0 and standard deviation 1) such that coefficients can be compared. Results are summarised in Table 3.6 and I detail each model below. In all models, gender and country fixed effects are included in all the models as control variables. In addition, I also collected the data about scholar's academic position (including four categories: student, post-doc, faculty and others) as a control variable in all the models. The past citations and the rank of the university of scholars are not available in my data set. However, it can be argued that the academic position can in general reflect the past citations of scholars.

Model 1 I start with Model 1 containing only standard brokerage proposed by Burt. As expected, standard brokerage is positively and significantly associated with the scholar's scientific impact ($r = 0.542, p < 0.001$).

Model 2-3 In Model 2, I only include gender-based brokerage measures, and results suggest that gender-based intra-brokerage is positively and significantly associated with the scholar's scientific impact ($r = 0.599, p < 0.001$) while the relationship between gender-based inter-brokerage and the scholar's scientific impact is not significant ($r = -0.257, p > 0.1$). Similar to Model 2, in Model 3, I only include country-based brokerage measures, and results show that country-based intra-brokerage is positively and significantly associated with the scholar's scientific impact ($r = 0.346, p < 0.1$) and country-based inter-brokerage is also positively and significantly associated with the scholar's scientific impact ($r =$

$0.388, p < 0.01$).

Model 4 Model 4 is my final model. I intend to use measures from both gender-based and country-based brokerage which are statistically significant in Models 2-3. Since the Pearson correlation coefficient between gender-based and country-based intra-broker is 0.918 indicating they are highly correlated, they cannot be both included in the model to avoid multicolinearity issue. In this case, I include gender-based intra-brokerage, country-based inter-brokerage and their interaction term. Both gender-based intra-brokerage ($r = 0.485, p < 0.1$) and country-based inter-brokerage ($r = 0.457, p < 0.05$) show positive and significant association with scholar's scientific impact, and thus results in Model 4 are consistent with Models 2-3. In addition, the interaction term of these two measures is also significant indicating that one brokerage measure can be moderated by the other. The interaction effect is also visualised in Figure 3.11. I further performed tests for multicollinearity (Mean VIF = 4.36) with "estat vif" command in Stata, omitted variables ($p = 0.3042$) with "estat ovtest, rhs" command in Stata and heteroskedasticity ($p = 0.9996$) with "estat imtest, white" in Stata, and results suggest that Model 4 meets these assumptions of OLS regression. Note that insignificant p-values imply a model that passes the tests for omitted variables and heteroskedasticity. I have also plotted scatter plots between residuals and independent variables. There are no discernible patterns in the plots suggesting that the model is not violating the zero mean conditional assumption.

I did not include the standard brokerage in my models because it highly correlates with gender-based and country-based intra-brokerage with Pearson correlation coefficients over 0.9. Adding the standard brokerage in the regression models causes multicollinearity issues. For example, the mean VIF will increase to 28.13 if the standard brokerage is included in Model 4.

Table 3.6: OLS estimates of the association between brokerage and scientific impact of a scholar.

	(1)	(2)	(3)	(4)
	ln(citations)	ln(citations)	ln(citations)	ln(citations)
Standard brokerage (Burt)	0.542^{***}			
	(0.144)			
Gender-based intra-brokerage		0.599^{***}		0.485^{+}
		(0.155)		(0.258)
Gender-based inter-brokerage		-0.257		
		(0.318)		
Country-based intra-brokerage			0.346^{+}	
			(0.182)	
Country-based inter-brokerage			0.388^{**}	0.457^{*}
			(0.135)	(0.196)
Gender-based intra-brokerage × Country-based inter-brokerage				-0.0863^{+}
				(0.0496)
Gender	Fixed	Fixed	Fixed	Fixed
Country	Fixed	Fixed	Fixed	Fixed
Position	Fixed	Fixed	Fixed	Fixed
Constant	9.511^{***}	9.438^{***}	9.025^{***}	8.829^{***}
	(0.626)	(0.640)	(0.807)	(0.898)
Number of samples	75	75	75	75
R^2	0.646	0.649	0.659	0.661

Robust standard errors in parentheses

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

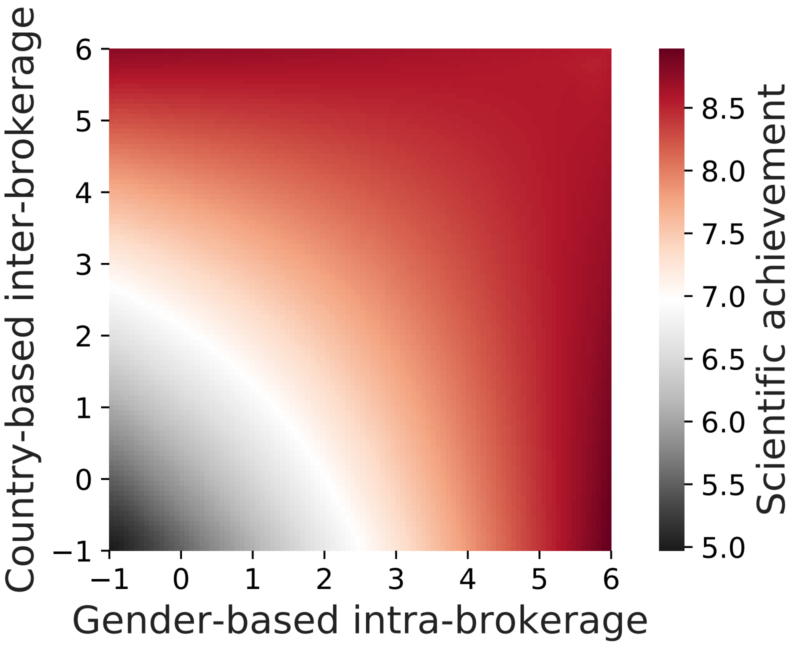


Figure 3.11: Visualisation of interaction effect of Model 4 in Table 3.6. This plots the scientific impact of a scholar predicted as a function of gender-based intra-brokerage (standardised) and country-based inter-brokerage (standardised).

3.6 Discussion and conclusion

In this chapter, I addressed the problem that simply considering only the network structure while ignoring the non-topological attributes of the actor might not provide a comprehensive perspective on the structural foundations of social capital. To address this issue, I have proposed intra- and inter-brokerage measures for quantifying open structures explicitly in terms of the non-topological attributes of the interacting actors. The relevant formalisations of intra- and inter-brokerage have been thoroughly developed in directed and weighted networks by closely following and extending the formalisation of network effective size proposed by Burt. Besides, in the case of undirected and unweighted networks, simplified formalisations of intra- and inter-brokerage measures have been suggested, and the relationship between these two measures and the intra- and inter-local clustering coefficients are derived as well.

As a case study, I applied three sets of brokerage measures (standard, intra- and inter-brokerage) on a co-authorship network, and showed that the intra- and inter-brokerage measures can capture distinct brokerage information compared with standard brokerage measures. Thus defining such intra- and inter-brokerage measures as a function of certain attributes of the actors can provide finer-grained perspectives on social capital. Moreover, I examined the association between the proposed measures and nodes' performance-based outcomes, which further improves our understanding of social capital and how actors can extract value from different types of open social structures.

3.6.1 Implications for research

My study offers a deeper perspective on Burt's brokerage and structural hole theory. Most studies in the past few years in the community of social network analysis have leveraged only the network structure to quantify the social capital. This has inevitably resulted in a number of measures of social capital that do not reflect the non-topological attributes of the nodes. However, given a fixed structure, an actor may well benefit from different brokerage opportunities to achieve better performance depending on the node's non-topological attributes (e.g., gender, race, education). My proposed intra- and inter-brokerage measures aim precisely to shed light on the salience of such non-topological attributes for social capital, and capture the nuances of an integrated perspective that would remain otherwise hidden by using the general brokerage measure proposed by Burt.

The proposed novel brokerage measures will open new doors to re-examine classic sociological questions related to the advantages of structural holes. Although the structural hole theory tells us that an actor in a social network enjoying higher brokerage opportunities can be better off, it does not provide insights about whether intra- or inter-brokerage based on a node attribute is needed. With the new measures proposed in this chapter, future studies can test the hypotheses related to how intra- and inter-brokerage based on a non-topological node attribute are associated with different performance-based outcomes in various social networks.

3.6.2 Implications for practice

As the node attributes in social networks are increasingly available in the era of big data (Qian et al., 2021b), there are unprecedented opportunities to combine network structure and non-topological node attributes to quantify the social capital of actors. Researchers could use the Python package I provided to easily measure the intra- and inter-brokerage based on a certain attribute of nodes. This will allow them to incorporate these two complementary perspectives to offer deeper insights on the role of social capital in the social sciences.

3.6.3 Limitations

In this chapter, I mainly considered a categorical attribute of nodes (e.g., gender) based on which an alter can naturally be classified as a member of either an intra-group or an inter-group associated with the focal node. However, many real-world attributes are not categorical but continuous (e.g., age). In this case, the proposed measures cannot be directly applied. However, one could convert the continuous variable into a discrete version, thus becoming a categorical variable. By so doing, my proposed measures can then be used with, and extended to, non-categorical data. It should also be noticed that my OLS regressions results cannot be interpreted as causal relationships. It would be interesting to leverage advanced causal inference methods, e.g., instrumental variable, to study whether the proposed intra- and inter-brokerage measures have causal effects on performance measures in social networks.

3.7 Contribution to the literature

Here I will summarise the main contributions of this chapter to extant literature on social capital. I developed a set of new network measures – "intra- and inter-brokerage" – that combine nodes' topological and non-topological features to extract sources of social capital in social networks. The proposed measures can be widely applied to weighted and directed networks as well as unweighted and undirected networks. I have addressed the long-debated problem in the literature that only considering network structure while ignoring non-topological node features cannot provide a comprehensive understanding of social capital (Aral and Van Alstyne, 2011; Fleming et al., 2007; Gould and Fernandez, 1989; Schilling and Fang, 2014; Shipilov and Li, 2008; Ter Wal et al., 2016; Uzzi, 1996). More specifically, my proposed brokerage measures can overcome this problem by extracting an actor's distinct brokerage types or roles, thus helping us gain a deeper understanding of the brokerage behaviour in social systems and offering the potential to study their relationships with an actor's performance. In addition, I also provide open-source code in Github that can allow researchers to easily apply my new measures and will offer opportunities to conduct new empirical studies on social capital.

Chapter 4

Network foundations of the scientific performance of cities

4.1 Introduction

Cities have long been regarded as the main engine of social innovation and wealth creation (Bettencourt et al., 2007a). For instance, Ref. (Li et al., 2017) suggests that cities worldwide have nowadays accounted for over 50% of the population, more than 80% of the world's wealth, and at least 90% of the innovation. As the main producers of scientific knowledge and innovation, scientists live and work primarily in cities (Vaccario et al., 2020; Verginer and Riccaboni, 2020a,b). In modern science, scientists usually do not work alone, but collaborate with others, which can potentially allow them to access interdisciplinary scientific expertise, produce more high-impact publications, and pursue new scientific interests (Evans et al., 2011; Lambiotte and Panzarasa, 2009; Pain, 2018; Qian et al., 2017). In particular, scientists collaborate not only with scientists in their own cities, but also with scientists in other cities. Indeed, international collaboration in science, an important modern phenomenon, has increased dramatically in the last two decades (Adams, 2013; Chen et al., 2019; Leydesdorff and Wagner, 2008; Scellato et al., 2015; Wagner and Leydesdorff, 2005).

The typical way to measure scientific collaboration relies on the fact that scientists co-author published papers together, which can be described with a special type of social networks – scientific collaboration networks (Newman, 2001a,c). In this case, two scientists are considered connected if they have co-authored one or more papers. Moreover, two scientists who are connected in scientific collaboration networks can be seen as scientific acquaintances because most people who have been co-authors usually know each other quite well (Newman, 2001d). Hence the network of scientific collaboration is a genuine social network of scientists.

In the social sciences, it is widely acknowledged that social capital, which can be extracted from social networks, plays an important role in maintaining or hindering a wide range of performance-related outcomes at the individual and group levels (Granovetter, 2005, 1973; Latora et al., 2013; Li et al., 2013). As cities are becoming central loci of scientific activities (Bettencourt, 2013), it is therefore essential to investigate the association between scientific network collaboration patterns and scientific performance at the city level. Although previous studies in recent years have explored scientific collaboration networks at the country or the affiliation level (Cantner and Rake, 2014; Graf and Kalthaus, 2018; Guan et al., 2016), it is surprising that there are very few studies conducted at the city level (Guan et al., 2015), considering the growing demand for more theoretical and empirical research on scientific collaboration at the city level (Neal, 2011). In addition, these very few studies concerned with the city level usually focus directly on the structural patterns of the inter-city collaboration networks, where nodes are cities, and links are simply constructed to reflect the scientific collaboration between geographical places. Thus the resulting inter-city collaboration network of cities aggregated from the individual scientist level lacks the actual collaboration patterns of scientists within and across cities.

To address this limitation, in this chapter I will propose to measure the social capital of a city by using the structural patterns of the scientific collaboration network of its internal scientists and external collaborators. Indeed, a city may be at the forefront of scientific performance not simply because it has elite resident scientists, but also as a result of the collaborative links that the scientists living in the city have with scientists in other places. To assess the interplay between collaboration patterns and scientific performance at the city level, it is therefore essential to account for both internal scientists and external collaborators (Hristova et al., 2016).

To this end, I collected the data from Medline which is one of the leading bibliometric data platforms, and is a publicly accessible repository of over 26 million publication records. Using two disambiguated data sets, Mapaffil (Torvik, 2015) and Author-ity (Torvik and Smalheiser, 2009), I tracked authors across publications, associated them to the city of their affiliation indicated in the publications, and established their collaboration ties. More specifically, this allowed me to create longitudinal global scientific collaboration networks of scientists using moving two-year time windows covering the period between 1990 and 2006. In each time window, each author is associated with her resident city. Therefore, for each city, its internal scientists and external collaborators were identified and their collaboration patterns were subsequently captured by the collaboration network through a few sets of network measures which reflect the social capital of a city from different angles (e.g., brokerage) that may be associated with scientific performance.

Furthermore, in each time window, I quantify the scientific performance of a city from two key perspectives: impact and innovation. More specifically, the scientific impact of a city in a time window was measured by the impact factors of the journals in which its internal scientists published papers during this window. Similarly, the scientific innovation of a city was measured by the number of new MeSH (medical subject headings) terms of papers, reflecting the new scientific knowledge. The impact factors of journals were collected from SCImago and the MeSH terms of papers were provided by Medline.

The remainder of this chapter is organised as follows. First, Section 4.2 presents the theoretical framework and hypotheses tested in the Chapter. Second, Section 4.3 describes the data, network approach, measures, and statistical models for this study. Third, Section 4.4 introduce descriptive and regression results obtained from my analysis. Fourth, Section 4.5 explores the significance of the findings of my work, discusses the policy implications of the results, outlines the limitations of this study and offers some avenues for further research. Finally, Section 4.6 summarises the findings and their contribution to the literature.

4.2 Theory and hypotheses

In this chapter, I propose that both internal scientists and external collaborators need to be considered. On the one hand, internal scientists are obviously essential because they are the main contributors to a city's scientific performance. On the other hand, external collaborators also need to be included due to the increasing number of inter-city collaborations between scientists (Adams, 2013). This is likely attributable to the development of modern communication technology that has reduced the cost of inter-city collaboration (Adams et al., 2005). The interactions of internal scientists and external collaborators of a city can be captured by the scientific collaboration network where nodes are scientists and links represent the co-authorship of scientific publications.

Recent studies have shown that central cities can benefit from their positions in inter-city scientific collaboration networks as they have more opportunities for knowledge creation and information diffusion (Guan et al., 2015). However, the inter-city scientific collaboration network is constructed by simply aggregating the collaboration between scientists in different cities, thus missing scientists' actual structural collaboration patterns. This approach might largely limit our understanding of the interplay between social capital and the scientific performance of cities. In this chapter, by drawing on the scientific collaboration network of internal scientists and external collaborators, the obtained integrated collaboration network can offer deeper insights into the social capital of a city.

4.2.1 Brokerage and scientific performance

Among various network mechanisms of social capital, I focus on open structures (Burt, 2009), rich in brokerage opportunities. The benefits that actors can extract from open structures have been thoroughly discussed in Chapter 2.

The brokerage of a city can be further divided into two finer-grained types – internal brokerage and external brokerage – by considering solely internal scientists and external collaborators, respectively. Internal brokerage focuses on internal residents of a city and looks at the absence of ties among these internal residents. By contrast, external brokerage focuses on a city's external collaborators and quantifies the absence of ties among these external collaborators. On the one hand, internal brokerage reflects the degree of non-redundant information that a city can extract from internal resident scientists. The scientific collaboration network of internal scientists of a city rich in structural holes can allow the city to possess information benefits and control benefits. On the contrary, a city associated with highly connected internal scientists with fewer structural holes might be limited by redundant ideas and information, becoming less impactful and innovative in terms of scientific performance. On the other hand, although external brokerage of a city describes the richness of structural holes among external collaborators, which is similar to internal brokerage, the two types of brokerage may have a distinct relationship with the scientific performance of a city. Meeting with scientists in the same affiliation or city to discuss and collaborate is relatively simple. By contrast, it may be more challenging to maintain collaboration with external collaborators due to time difference and geographic distance. In this case, higher external brokerage, indicating a less cohesive structure between scientists and their external collaborators, may result in the lack of a sense of belonging (Coleman, 1988), trust (Coleman, 1994; Reagans and McEvily, 2003; Uzzi, 1997) and more costly communication channels to exchange complex and proprietary information (Hansen, 1999; Uzzi, 1997). This line of reasoning resonates with the theory of "closed" structure of social capital advocated by Coleman (Coleman, 1988). Based on the above arguments and discussion, I propose the following hypotheses:

Hypothesis 1(a-b). The internal brokerage of a city is positively associated with its scientific impact (1(a)) and innovation (1(b)).

Hypothesis 1(c-d). The external brokerage of a city is negatively associated with its scientific impact (1(c)) and innovation (1(d)).

4.2.2 Strong ties and scientific performance

Another fundamental network mechanism of social capital is related to the concept of "strength of weak ties". In a seminal paper (Granovetter, 1973), Granovetter defined the strength of a tie as "a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterise the tie" (p. 1361). In addition, Granovetter argued that the most useful network contacts are through "weak ties". Intuitively, this is because weak ties can allow individuals to be connected with a more diverse set of alters such that they can increase the ranges of their networks (Uzzi, 1996; Uzzi and Spiro, 2005). A lot of work has suggested that unusual and fruitful recombination of existing components (e.g., ideas, information, and devices) is an important source of innovation, and access to diverse pools of knowledge and perspectives can increase the chance of creating novel ideas (Bettencourt et al., 2007b; Muller and Zenker, 2001; Singh, 2008). On the contrary, strong ties allow individuals to bind themselves to each other, which in turn makes them more redundant in obtaining new information. In my context concerned with a city, as the proportion of strong ties among internal scientists increases, it is more likely that scientists connected by strong ties become more similar to each other, which likely prevents them from producing more impactful and novel work. In addition, as scientists have limited time and energy, maintaining strong ties is costly, and thus having too many strong relationships can be inefficient in the long run (McFadyen and Cannella Jr, 2004; Perry-Smith and Shalley, 2003; Wang, 2016). On the other hand, strong ties among external collaborators may play a different role in shaping scientific performance. As collaboration among internal and external partners may be hindered by time differences and geographical distances, a city's scientific performance is likely to benefit from the strong ties connecting its external collaborators, which in turn may exert mitigating effects on the ease of collaboration. In this sense, the reduced intellectual diversity associated with stronger bonds between geographically distant collaborators is likely to be more than compensated for by the benefits in terms of lower coordination and communication costs. Therefore, I hypothesise that:

Hypothesis 2(a-b). The proportion of internal strong ties of a city is negatively associated with its scientific impact (2(a)) and innovation (2(b)).

Hypothesis 2(c-d). The proportion of external strong ties of a city is positively associated with its scientific impact (2(c)) and innovation (2(d)).

4.2.3 Diversity and scientific performance

Diversity, the new orthodoxy in city planning, has long been considered important because it not only makes cities more appealing but also is the catalyst of scientific performance (AlShebli et al., 2018), economic productivity (Florida, 2002; Jacobs, 1985, 2016), and social justice (Young, 2011). Since the 1960s, researchers from different disciplines have been devising strategies for urban redevelopment which stimulate both physical and social diversity (Fainstein, 2005), thus suggesting that diversity is a critical characteristic of cities. This may partly be explained by the fact that cities with diverse knowledge, experience, and skills among their internal scientists and external collaborators can benefit from the integration of expertise, successful project implementation, and accelerated cycle time for new knowledge creation (Cummings, 2004; Eisenhardt and Tabrizi, 1995). In addition, diversity is a complex concept since cities can be diverse in terms of various properties, such as ethnicity, gender, age, and socio-economic background (Fainstein, 2005). Indeed, depending on the research fields and questions, one of these attributes or a mix of them may be applied by researchers in different studies. This chapter focuses on the geographical diversity of a city as a function of the residents' external collaborators. A city with high diversity in terms of geographical places of external collaborators may enjoy diverse knowledge and expertise from different backgrounds and research cultures (Barjak and Robinson, 2008). Previous studies have suggested that geographically diverse scientific collaboration is associated with high research impact at the university and country levels (Abbasi and Jaafari, 2013). However, little attention has been paid to this association at the city level. Based on the above arguments and discussion, I propose the following hypotheses:

Hypothesis 3(a-b). The geographical diversity of a city is positively associated with its scientific impact (3(a)) and innovation (3(b)).

4.2.4 Interaction effects

Furthermore, I explore the idea that brokerage, the proportion of strong ties, and geographical diversity may not play their roles independently when considering their relationship with scientific performance. Specifically, I focus on two types of interactions in this chapter: (i) the interaction between brokerage and proportion of strong ties; and (ii) the interaction between brokerage and geographical diversity. On the one hand, the salience of brokerage for performance may be mediated by the proportion of strong ties as a result of the collaboration structure rich in trust and common knowledge base brought by strong ties, which can provide advantages for knowledge transfer (Hansen, 1999; Reagans and McEvily, 2003) and knowledge creation (McFadyen and Cannella Jr, 2004; McFadyen et al., 2009). On the other hand, the role of brokerage may also be amplified by geographical diversity. For example, a city may benefit more from an increase in brokerage due to the diverse knowledge and information brought by high diversity (Fleming et al., 2007; Østergaard et al., 2011). Therefore, I can propose the following hypotheses:

Hypothesis 4(a-b). The proportion of internal strong ties of a city moderates the relationship between its internal brokerage and scientific impact (4(a)) and innovation (4(b)).

Hypothesis 4(c-d). The proportion of external strong ties of a city moderates the relationship between its external brokerage and scientific impact (4(c)) and innovation (4(d)).

Hypothesis 5(a-b). The geographical diversity of a city moderates the relationship between its internal brokerage and scientific impact (5(a)) and innovation (5(b)).

Hypothesis 5(c-d). The geographical diversity of a city moderates the relationship between its external brokerage and scientific impact (5(c)) and innovation (5(d)).

4.3 Material and methods

4.3.1 Data

I combined four large-scale data sets for my analysis: Medline, Author-ity, MapAffil, and SCImago. In what follows I shall detail each of them in turn.

First, Medline[1] is a publicly accessible repository of over 26 million publication records mostly related to life sciences. In Medline, the earliest publication year dates back to 1867. Here, I will focus on the period between 1990 and 2006

[1] ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline-2018-sample/

because the other two high-quality disambiguation data sets of scientists (Author-ity (Torvik and Smalheiser, 2009)) and affiliations (MapAffil (Torvik, 2015)) are restricted to this time period. Note that each Medline publication is associated with a set of MeSH terms assigned to describe the content of the publication. Second, Author-ity, developed by Ref. (Torvik and Smalheiser, 2009), contains around 9 million unique scientists disambiguated from over 61 million scientists' names that appeared in the Medline publications. Third, MapAffil[2], developed by Ref. (Torvik, 2015), associates over 37 million scientists' affiliations that appear in the Medline publications with disambiguated cities. This data set allows me to map the affiliation string to the city in which this affiliation is located. By merging Medline and Author-ity, I can obtain the necessary data to uniquely identify a scientist across publications, which subsequently allows me to construct global scientific collaboration networks. By further merging MapAffil with the previous two data sets, I can extract a scientist's publication history with the geographical location at the city level, which can be used to map the scientist to his or her resident city. The last mapping step is necessary because otherwise the scientist's affiliation listed in the publication would not be disambiguated, and distinct versions of "Boston University" would exist in the data set. Notice that from Mapaffil the locations of scientists are either at high-resolution (e.g., "Bethnal Green, London, UK") or low-resolution level (e.g., "London, UK") of a city. I mapped the high-resolution locations to the low-resolution level as suggested in Refs. (Verginer and Riccaboni, 2020a,b). For example, "Bethnal Green, London, UK" would be mapped to "London, UK". Finally, SCImago[3] provides me with

[2] http://abel.lis.illinois.edu/cgi-bin/mapaffil/search.pl

[3] https://www.SCImagojr.com/

access to yearly impact factor scores for a large portion of journals indexed in Medline. I shall detail how I calculate the scientific impact of cities based on SCImago impact factors in Section 4.3.3.

4.3.2 Network construction

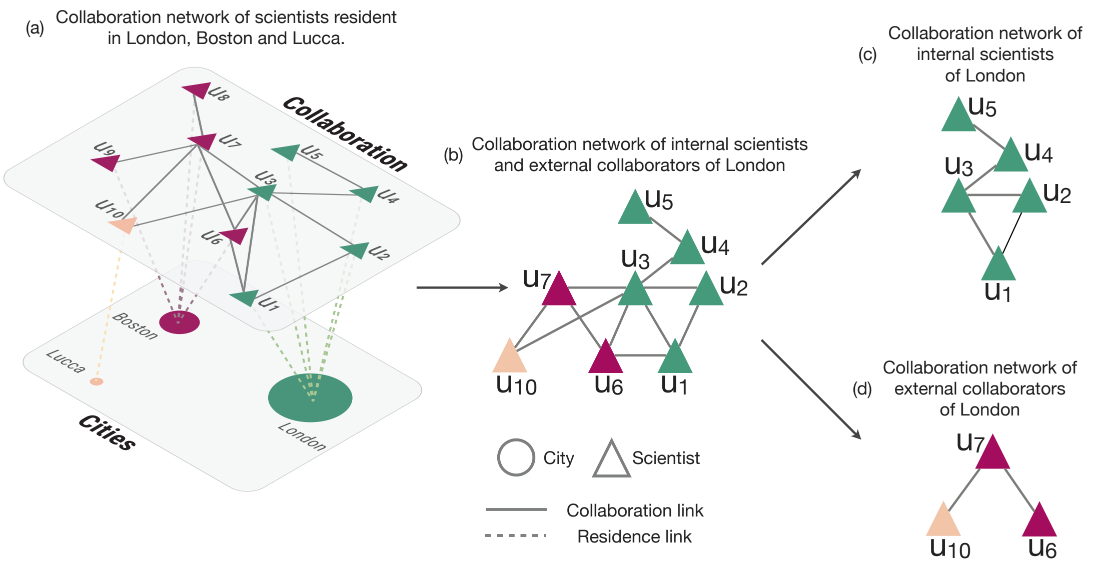

(a) Collaboration network of scientists resident in London, Boston and Lucca.
Collaboration
Cities
u8
u7
u9
u10
u5
u3
u4
u6
u2
u1
Boston
Lucca
London
(b) Collaboration network of internal scientists and external collaborators of London
u5
u4
u3
u7
u2
u10
u6
u1
City
Scientist
Collaboration link
Residence link
(c) Collaboration network of internal scientists of London
u5
u4
u3
u2
u1
(d) Collaboration network of external collaborators of London
u7
u10
u6

Figure 4.1: Schematic diagram of the workflow for constructing the collaboration networks of internal scientists and external scientists of a city.

Given a focal window $\mathcal{T}_t = [t, t + \tau)$ where t represents the focal year and τ represents the number of years in each window, I can construct a scientific collaboration network where nodes are scientists and links refer to co-authorship of papers. For the length of the window $\mathcal{T}_t$, I set it at 2 years, i.e., $\tau = 2$. In the following sections, I will focus on one time window $\mathcal{T}_t$ and thus omit the t symbol in the notations when there is no ambiguity. In each time window, I construct the scientific collaboration network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes $u \in \mathcal{V}$ are scientists, and undirected weighted links, denoted by $\mathcal{E}$, represent (the intensity of) co-authorship between scientists (see Figure 4.1 (a)). Note that $\mathcal{V}$ is composed of scientists who have publications in the focal window $\mathcal{T}_t$. I assign weights $w_{\mathrm{S}}(u, v)$ to links based on the number of papers two scientists u and v co-authored in $\mathcal{T}_t$ and the number of authors in each paper, as suggested by Newman in Ref. (Newman, 2001c).

Furthermore each scientist u is associated with a unique city of residence $L(u)$ in time window $\mathcal{T}_t$. To determine $L(u)$ in $\mathcal{T}_t$, I choose the longest uninterrupted sequence of cities closest to t in the time window, following the rules suggested by authors of Refs. (Verginer and Riccaboni, 2020a,b). I choose this method since it can discard ambiguous affiliations in publication sequences with spurious affiliations, such as multiple affiliations in the same year, but either of these appears only once.

I construct the collaboration network of city i $\mathcal{G}(i)$ consisted of its internal scientists and external collaborators (see Figure 4.1 (b)). I denote the set of both internal scientists and external collaborators as $\mathcal{N}_{\mathrm{S}}(i)$. I further extract the collaboration network $\mathcal{G}^{\mathrm{in}}(i)$ of its internal scientists (see Figure 4.1 (c)) and the collaboration network $\mathcal{G}^{\mathrm{ex}}(i)$ of its external collaborators (see Figure 4.1 (d)). I denote the set of internal scientists of city i as $\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)$, and the set of external collaborators of city i as $\mathcal{N}_{\mathrm{S}}^{\mathrm{ex}}(i)$. In this way, the extracted networks will allow me to define measures (see Section 4.3.3) to properly test my hypotheses (see Section 4.2).

4.3.3 Measures

In what follows, I will introduce the dependent variables, control variables, and independent variables. The main notations used in this chapter are summarised in Table 4.2.

Dependent variables

I measure the scientific performance of city i in $\mathcal{T}_t$ from two perspectives: scientific impact, $\Gamma(i)$, and scientific innovation, $\Delta(i)$. I quantify $\Gamma(i)$ and $\Delta(i)$ in three steps: (i) Let me denote the set of papers published in $\mathcal{T}_t$ as $\mathcal{P}$. For each paper $p \in \mathcal{P}$, I first obtain its scientific impact, $\gamma(p)$, and scientific innovation, $\delta(p)$. $\gamma(p)$ is measured by the 2-year impact factor of the journal where the paper was published. $\delta(p)$ is measured by the number of new MeSH terms associated with paper p. A MeSH term is considered as new for up to two years since its first appearance in Medline; (ii) For paper p, I calculate the fraction, $\alpha(i, p)$, of city i representing the ratio $\alpha(p, i)$ between the number of authors of paper p that are resident in i and the total number of authors of paper p. For instance, if paper p is co-authored by 4 scientists, three of whom reside in London and the other in Boston, we would have: $\alpha(p, \text{London}) = 0.75$, and $\alpha(p, \text{Boston}) = 0.25$. This measure allows me to proportionally allocate the scientific performance of each paper to various relevant cities. If the impact factor of the journal of paper p is 4.5 and it contains 3 new MeSH terms, i.e., $\gamma(p) = 4.5$ and $\delta(p) = 3$, London will be associated with a value of impact $\gamma(p) \times \alpha(p, \text{London}) = 4.5 \times 0.75 = 3.375$ and a value of innovation $\delta(p) \times \alpha(p, \text{London}) = 3 \times 0.75 = 2.25$. Similarly, Boston will be associated with a value of impact $\gamma(p) \times \alpha(p, \text{Boston}) = 4.5 \times 0.25 = 1.125$ and a value of innovation $\delta(p) \times \alpha(p, \text{London}) = 3 \times 0.25 = 0.75$; and (iii) For each paper in $\mathcal{P}$, I iterate step (ii) above and proportionally assign its scientific impact and innovation to the corresponding cities. Then, for each city i, I aggregate the scientific impact and innovation assigned by all the papers in $\mathcal{P}$ to obtain the final scientific impact $\Gamma(i)$ and innovation $\Delta(i)$ for city i in $\mathcal{T}_t$. Formally, $\Gamma(l)$ and $\Delta(l)$

can be expressed as:

$$\begin{aligned}\Gamma(i) &= \sum_{p\in\mathcal{P}(\mathcal{T}_t)} \gamma(p) \times \alpha(p,i),\\ \Delta(i) &= \sum_{p\in\mathcal{P}(\mathcal{T}_t)} \delta(p) \times \alpha(p,i).\end{aligned} \tag{4.1}$$

4.3.4 Control variables

Size

To control for scale influence, I take the logarithm of the total number of internal scientists and external collaborators as the size of the city i. I denote the size of city i as $Q(i)$ such that

$$Q(i) = ln(|\mathcal{N}_{\mathrm{S}}(i)| + 1) \tag{4.2}$$

where $|\cdot|$ represents the number of elements in a set.

Centralisation

To control for the heterogeneity of degree distribution of the collaboration network of internal scientists and external collaborators of city i, I select centralisation (Freeman, 1978) as the control variable. In my case, I define the centralisation $H(i)$ of city i in two steps: (i) by calculating the sum of the degree difference between the highest degree node and all other nodes in $\mathcal{N}_{\mathrm{S}}(i)$; and (ii) by normalising the value of (i) by the maximum value of such sum of degree difference in any network containing $|\mathcal{N}_{\mathrm{S}}(i)|$ nodes. Note that the maximum centralisation of a network with a fixed number of nodes can be achieved by a star network where one central node connects all other nodes, and there are no connections between all other nodes. Formally, I define $H(i)$ as:

$$H(i) = \frac{\sum\limits_{u \in \mathcal{N}_{\mathrm{S}}(i)} (\deg(u^*) - \deg(u))}{(|\mathcal{N}_{\mathrm{S}}(i)| - 1)(|\mathcal{N}_{\mathrm{S}}(i)| - 2)}, \tag{4.3}$$

where $\deg(u)$ is the unweighted degree (i.e., number of collaborators) of scientist u in $\mathcal{N}_{\mathrm{S}}(i)$. u^* is the scientist with the highest degree in $\mathcal{N}_{\mathrm{S}}(i)$. $|\mathcal{N}_{\mathrm{S}}(i)| - 2$ is the degree difference between the central node and any other surrounding node in a star network containing $|\mathcal{N}_{\mathrm{S}}(i)|$ nodes.

Betweenness and closeness centrality

Recent work has suggested that the scientific performance of a city is influenced by its position in the inter-city collaboration network where nodes are cities and a link is established between two cities if scientists in these two cities have co-authored scientific publications (Guan et al., 2015). The inter-city collaboration network (see Figure 4.2) can be seen as an aggregation of the collaboration network of scientists studied in my work. In this case, the strength $w_{\mathrm{L}}(i, j)$ of the link between cities i and j in the inter-city collaboration network is obtained by aggregating the weights of collaboration of scientists from these two cities.

To control for the position of city i in the inter-city collaboration network, I consider two traditional centrality measures: betweenness $BC(i)$ and closeness $CC(i)$ centralities (Freeman, 1977). On the one hand, betweenness centrality measures the potential of gatekeeping, brokering and controlling the information flow and the ability to liaise between otherwise separate parts of the network.

A node with higher betweenness centrality that lies on communication paths can control the communication flow between others, and thus may well play an important role for the functioning of the network. On the other hand, closeness centrality measures, for a given node, the expected time until the arrival of whatever is flowing through the network. A node with higher closeness centrality is considered important as it is close to most other nodes. In my context, a city with high betweenness and closeness centralities likely plays the role of a gatekeeper or broker of knowledge, and it can also diffuse or receive knowledge from others in a relatively short time. I use NetworkX[4] to calculate these two weighted centrality measures. More specifically, as these two measures are based on the shortest paths between nodes, I use the reciprocal of the weight of collaboration since a higher weight of collaboration of two cities suggests that the distance between them is shorter.

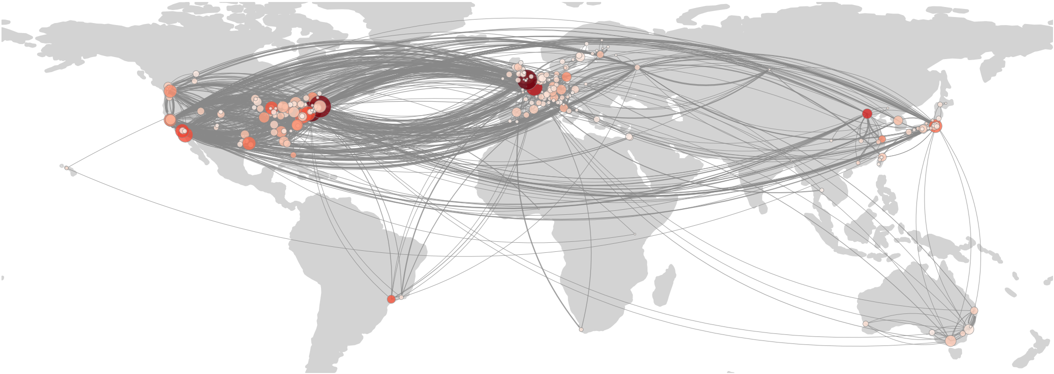


Figure 4.2: **Inter-city collaboration network for focal year** 2006**.** The figure shows the largest connected component of the inter-city collaboration network. The size of a node is proportional to the city's scientific impact. The colour intensity of a node is proportional to its scientific innovation. The width of an edge is proportional to its weight. Edges with weights less than 20 are removed. Self-loops are also excluded.

[4] https://networkx.github.io/

4.3.5 Independent variables

Brokerage

The “geo-social brokerage potential” of a city refers to the opportunities of brokerage the city can offer with respect to the collaboration networks of its residents and those connected to them (Hristova et al., 2016). Following the formalisation proposed in Refs. (Borgatti, 1997; Burt, 2009; Hristova et al., 2016), I first measured the internal brokerage $S^{\mathrm{in}}(i)$ of city i in two steps: (i) by calculating the non-redundant portion of the the collaboration network of the city’s internal scientists; and (ii) by normalising the value of (i) by $|\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)|$, resulting in the fraction of non-redundant contacts of city i’s collaboration network of internal scientists. Formally, $S^{\mathrm{in}}(i)$ can be expressed as:

$$S^{\mathrm{in}}(i) = \frac{|\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)| - \frac{\sum\limits_{u,v \in \mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)} e(u,v)}{|\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)|}}{|\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)|} = 1 - \frac{\sum\limits_{u,v \in \mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)} e(u,v)}{|\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)|^2}, \tag{4.4}$$

where $e(u,v)$ is 1 if two nodes u, v are connected and 0 otherwise.

Similarly, I further measure the external brokerage $S^{\mathrm{ex}}(i)$ of city i by only considering the collaboration network of the city’s external collaborators. Formally, $S^{\mathrm{ex}}(i)$ can be expressed as:

$$S^{\mathrm{ex}}(i) = 1 - \frac{\sum\limits_{u,v \in \mathcal{N}_{\mathrm{S}}^{\mathrm{ex}}(i)} e(u,v)}{|\mathcal{N}_{\mathrm{S}}^{\mathrm{ex}}(i)|^2}. \tag{4.5}$$

Proportion of strong ties

The strength of a tie between two scientists can be measured based on the number of papers they coauthored. In this case, I quantify the ratio between a city's internal strong ties and the total number of ties between the city's internal scientists:

$$R^{\text{in}}(i) = \frac{\sum_{u,v \in \mathcal{N}_{\text{S}}^{\text{in}}(i)} e^{\#}(u,v)}{\sum_{u,v \in \mathcal{N}_{\text{S}}^{\text{in}}(i)} e(u,v)}. \tag{4.6}$$

where $e^{\#}(u,v)$ is 1 if the tie between scientists u, v is a strong tie and 0 otherwise. Specifically, a tie is regarded as strong if two scientists co-author at least 4 papers in time window $\mathcal{T}_t$, i.e., they co-authored on average 2 papers each year. Similarly, the proportion of external strong ties can be defined as:

$$R^{\text{ex}}(i) = \frac{\sum_{u,v \in \mathcal{N}_{\text{S}}^{\text{ex}}(i)} e^{\#}(u,v)}{\sum_{u,v \in \mathcal{N}_{\text{S}}^{\text{ex}}(i)} e(u,v)}. \tag{4.7}$$

Geographical diversity

I define the geographical diversity $D(i)$ of city i in terms of the diversity of the strength of collaboration between city i and the cities associated with scientists in $\mathcal{N}_{\text{S}}(i)$.

To this end, I define the set of resident cities of the external collaborators of city i as $\mathcal{N}_{\text{L}}(i)$. I use Gini impurity to compute the geographical diversity $D(i)$ of $\mathcal{N}_{\text{L}}(i)$:

$$D(i) = 1 - \sum_{j \in \mathcal{N}_{\text{L}}(i)} P(i,j)^2, \tag{4.8}$$

where $P(i, j)$ is the probability of city i to have a collaboration with city j included in $\mathcal{N}_{\mathrm{L}}(i)$:

$$P(i, j) = \frac{w_{\mathrm{L}}(i, j)}{\sum\limits_{k \in \mathcal{N}_{\mathrm{L}}(i)} w_{\mathrm{L}}(i, k)}, \tag{4.9}$$

To illustrate how these measures are calculated, in Table 4.1 I focus on London and compute the corresponding measures based on the collaboration networks constructed according to Figure 4.1.

4.3.6 Statistical models

I estimate linear mixed-effects models (LMEMs), also referred to as multilevel models or hierarchical linear models. The LMEM (Searle et al., 2009) contains both fixed effects and random effects. LMEMs are particularly useful when data are organised into more than one level (i.e., nested data). In my case, as cities are nested within countries, I use the `mixed` module in Stata to fit a three-level mixed model with random intercepts at both the country and city levels. Thus my model has two random-effects. The first is a random intercept at the city level (level 2), and the second is a random intercept at the country level (level 3).

For either scientific impact or innovation, I will estimate three models: (i) a baseline model that includes only control variables (see Equation (4.10)); (ii) a model that includes all independent (control and independent) variables (see Equation (4.11)); and (iii) a model in which I also add interaction terms between independent variables (see Equation (4.12)). Formally, we have:

$$\begin{aligned}\text{Scientific performance}_{tij} &= \beta_0 + \beta_1 Q_{tij} + \beta_2 H_{tij} + \beta_3 BC_{tij} + \beta_4 CC_{tij} + \\ &\sum_{T=0}^{15} d_{T,t,i,j}\delta_T + \mu_{ij}^{(2)} + \mu_j^{(3)} + \epsilon_{tij} \qquad (4.10)\end{aligned}$$

$$\begin{aligned}\text{Scientific performance}_{tij} &= \beta_0 + \beta_1 Q_{tij} + \beta_2 H_{tij} + \beta_3 BC_{tij} + \beta_4 CC_{tij} + \\ &\beta_5 S_{tij}^{\text{in}} + \beta_6 S_{tij}^{\text{ex}} + \beta_7 R_{tij}^{\text{in}} + \beta_8 R_{tij}^{\text{ex}} + \beta_9 D_{tij} + \\ &\sum_{T=0}^{15} d_{T,t,i,j}\delta_T + \mu_{ij}^{(2)} + \mu_j^{(3)} + \epsilon_{tij} \qquad (4.11)\end{aligned}$$

$$\begin{aligned}\text{Scientific performance}_{tij} &= \beta_0 + \beta_1 Q_{tij} + \beta_2 H_{tij} + \beta_3 BC_{tij} + \beta_4 CC_{tij} + \\ &\beta_5 S_{tij}^{\text{in}} + \beta_6 S_{tij}^{\text{ex}} + \beta_7 R_{tij}^{\text{in}} + \beta_8 R_{tij}^{\text{ex}} + \beta_9 D_{tij} + \\ &\beta_{10}(S_{tij}^{\text{in}} \times R_{tij}^{\text{in}}) + \beta_{11}(S_{tij}^{\text{ex}} \times R_{tij}^{\text{ex}}) + \beta_{12}(S_{tij}^{\text{in}} \times D_{tij}) + \beta_{13}(S_{tij}^{\text{ex}} \times D_{tij}) + \\ &\sum_{T=0}^{15} d_{T,t,i,j}\delta_T + \mu_{ij}^{(2)} + \mu_j^{(3)} + \epsilon_{tij} \qquad (4.12)\end{aligned}$$

where t represents the focal year (level 1), i represents the city (level 2), and j represents the country (level 3). β_0 to β_{13} are fixed parameters. $\sum_{T=0}^{15} d_{T,t,i,j}\delta_T$ is the year fixed effect where $d_{T,t,i,j}$ is the dummy variable for T-th focal year. As there are in total 17 focal years, I have 16 dummy variables here. $\mu_{ij}^{(2)}$ is the level-2 (i.e., city-level) random intercept, $\mu_j^{(3)}$ is the level-3 (i.e., country-level) random intercept, ϵ_{tij} is the level-1 (i.e., occasion-level) error term. Scientific performance will be measured either as scientific impact (Γ) or as scientific innovation (Δ).

Here I shall also introduce the assumptions underlying the above models: (i) The country-level random intercept ($\mu_j^{(3)}$) has zero expectation, given the independent variables; (ii) Similarly, the city-level random intercept ($\mu_{ij}^{(2)}$) has zero expectation, given the independent variables and $\mu_j^{(3)}$; (iii) There is zero correlation between independent variables and the random intercept at the country level (i.e., level-3 exogeneity); (iv) Similarly, there is zero correlation between independent variables and the random intercept at the city level (i.e., level-2 exogeneity); (v) There is zero correlation between random intercepts ($\mu_{ij}^{(2)}$ and $\mu_j^{(3)}$) across countries and cities; (vi) The variance of the random intercept at the country level is homoskedastic given the independent variables; (vii) Similarly, the variance of the random intercept at the city level is homoskedastic given the independent variables and $\mu_j^{(3)}$; (viii) The level-1 error term ϵ_{tij} has zero expectation, given the independent variables and the random intercepts ($\mu_j^{(3)}$ and $\mu_{ij}^{(2)}$); and (ix) There is zero correlation between the independent variables and the level-1 residual (i.e., level-1 exogeneity), and zero correlation between the level-1 residual and both random intercepts ($\mu_j^{(3)}$ and $\mu_{ij}^{(2)}$).

Unlike the two random intercepts, the level-1 residuals were not assumed to be homoskedastic. To this end, I used Stata to carry out the analysis with the above models and added the `residuals(independent, by(t))` to fit the models with heteroskedastic level-1 residuals over occasions (focal year, t). I also standardised (i.e., subtract the mean and divide by the standard deviation) all the independent and control variables to make the interpretations of coefficients more reasonable, considering the range of some covariates is very small. Moreover, I estimated robust standard errors by using `vce(robust)` in Stata.

4.4 Results

Table 4.3: Descriptive statistics

	Mean	SD	Min	Max	Correlation										
					1	2	3	4	5	6	7	8	9	10	11
Control variables															
1 Size, Q	6.570	1.166	4.094	10.532	1.000										
2 Centralisation, H	0.071	0.061	0.003	0.539	-0.131***	1.000									
3 Betweenness, BC	0.001	0.004	0.000	0.104	0.373***	-0.084***	1.000								
4 Closeness, CC	0.954	0.144	0.390	1.283	0.638***	-0.132***	0.195***	1.000							
Independent variables															
5 Internal brokerage, S^{in}	0.995	0.010	0.706	1.000	0.255***	-0.356***	0.061***	0.100***	1.000						
6 External brokerage, S^{ex}	0.931	0.082	0.109	0.999	0.402***	-0.538***	0.118***	0.489***	0.142***	1.000					
7 Proportion of internal strong ties, R^{in}	0.016	0.057	0.000	0.972	0.164***	0.260***	0.081***	0.144***	-0.379***	-0.040***	1.000				
8 Proportion of external strong ties, R^{ex}	0.074	0.176	0.000	1.000	0.356***	0.501***	0.129***	0.120***	-0.021**	-0.235***	0.314***	1.000			
9 Geographical diversity, D	0.930	0.075	0.153	0.993	0.592***	-0.065***	0.125***	0.314***	0.141***	0.448***	0.056***	0.153***	1.000		
Dependent variables															
10 Scientific impact, Γ	1548.287	3177.206	23.384	42328.692	0.652***	-0.185***	0.784***	0.410***	0.162***	0.262***	0.043***	0.181***	0.275***	1.000	
11 Scientific innovation, Δ	4.180	10.698	0.000	234.480	0.477***	-0.215***	0.588***	0.405***	0.131***	0.224***	0.023**	0.034***	0.205***	0.792***	1.000
Observations	10897														

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.4: Top 10 cities for focal year 2006. Non-US cities are in bold.

Ranking	Size, Q	Centralisation, H	Betweenness centrality, BC	Closeness centrality, CC	Internal brokerage, S^{in}	External brokerage, S^{ex}
1	**London, UK**	**Victoria, BC, Canada**	**Paris, France**	**Paris, France**	**Sao Paulo, Brazil**	Boston, MA, USA
2	Boston, MA, USA	Arlington, TX, USA	Boston, MA, USA	**London, UK**	**London, UK**	Bethesda, MD, USA
3	**Paris, France**	**Lancaster, Lancashire, UK**	**London, UK**	Boston, MA, USA	**Tokyo, Japan**	Birmingham, AL, USA
4	New York, NY, USA	Duluth, MN, USA	New York, NY, USA	Cambridge, MA, USA	New York, NY, USA	**Toronto, ON, Canada**
5	**Tokyo, Japan**	Eugene, OR, USA	Stanford, CA, USA	Stanford, CA, USA	**Guangzhou, China**	**Berlin, Germany**
6	Bethesda, MD, USA	South Bend, IN, USA	Cambridge, MA, USA	New York, NY, USA	**Paris, France**	Rochester, MN, USA
7	**Beijing, China**	Medford, MA, USA	**Moskva, Russia**	**Cambridge, Cambridgeshire, UK**	**Shanghai, China**	**Leiden, Zuid-Holland, Netherlands**
8	Baltimore, MD, USA	Fairfax, VA, USA	Bethesda, MD, USA	**Roma, Lazio, Italy**	Boston, MA, USA	**Basel, Switzerland**
9	Los Angeles, CA, USA	**Essex, UK**	**Tokyo, Japan**	Bethesda, MD, USA	**Madrid, Spain**	Indianapolis, IN, USA
10	San Diego, CA, USA	Kent, OH, USA	**Beijing, China**	**Oxford, Oxfordshire, UK**	**Sydney, NSW, Australia**	Portland, OR, USA

Ranking	Proportion of internal strong ties, R^{in}	Proportion of external strong ties, R^{ex}	Geographical diversity, D	Scientific impact, Γ	Scientific innovation, Δ
1	DeKalb, IL, USA	**Shigenobu-cho, Toon, Ehime, Japan**	**Paris, France**	Boston, MA, USA	**London, UK**
2	Upton, NY, USA	**Egham, Surrey, UK**	**London, UK**	New York, NY, USA	New York, NY, USA
3	Arlington, TX, USA	**Rostock, Germany**	Tucson, AZ, USA	**London, UK**	Boston, MA, USA
4	Stony Brook, NY, USA	**Perugia, Umbria, Italy**	**München, Germany**	**Paris, France**	**Paris, France**
5	Stanford, CA, USA	**Bochum, Germany**	**Wien, Austria**	San Diego, CA, USA	**Beijing, China**
6	**Pisa, Toscana, Italy**	**Bergen, Norway**	**Moskva, Russia**	Bethesda, MD, USA	Los Angeles, CA, USA
7	**Beijing, China**	Albany, NY, USA	New York, NY, USA	Baltimore, MD, USA	Bethesda, MD, USA
8	**Karlsruhe, Germany**	Louisville, KY, USA	State College, PA, USA	Philadelphia, PA, USA	San Diego, CA, USA
9	Norfolk, VA, USA	**Coventry, West Midlands, UK**	**Ciudad de Mexico, DF, Mexico**	Los Angeles, CA, USA	Philadelphia, PA, USA
10	South Bend, IN, USA	**Victoria, BC, Canada**	San Diego, CA, USA	Houston, TX, USA	**Sao Paulo, Brazil**

4.4.1 Descriptive statistics

I first constructed the collaboration networks of scientists for 17 time windows $\mathcal{T}_t$ with $t \in \{1990, 1991, ..., 2006\}$, respectively. In each $\mathcal{T}_t$, each scientist is associated with her resident city. Then I computed the measures proposed in Section 4.3.3, which results in a 17-year panel data set. In what follows, I will focus on the cities whose internal size ($|\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)|$) is at least 50 in all the 17 windows. The final data set is thus a balanced panel data which contains 641 cities grouped into 64 countries with focal years in the period $1990 - 2006$.

Table 4.3 shows the unstandardised descriptive statistics (mean, standard deviation, minimum, maximum, and pairwise Pearson correlation) included in my study. First, there is no strong correlation between control variables and independent variables, which mitigates the potential multicollinearity issue. Second, there is a strong positive correlation between scientific impact and innovation, which suggests that a city with higher scientific impact at the same time is likely to produce higher scientific innovation. Third, all control variables and independent variables (except centralisation) are positively correlated with scientific impact and innovation. In particular, size shows a very strong correlation with scientific impact and innovation. Last, my proposed measures can naturally provide the rankings of cities from different perspectives. As an example, I show the top 10 cities for the last focal year (i.e., 2006) in Table 4.4.

Table 4.5: Three-level hierarchical random-intercept model with heteroskedastic level-1 residuals over occasions (focal year). Independent and control variable are standardised by subtracting mean and dividing by standard deviation.

	(1)	(2)	(3)	(4)	(5)	(6)
	Scientific impact, Γ			Scientific innovation, Δ		
Internal brokerage, S^{in}		13.85***	13.68***		0.145***	0.183***
		(2.786)	(4.327)		(0.0200)	(0.0441)
External brokerage, S^{ex}		21.55***	12.77***		-0.175***	-0.0636
		(3.982)	(4.278)		(0.0619)	(0.0413)
Proportion of internal strong ties, R^{in}		-5.874	-3.320		0.0706**	0.0582*
		(6.009)	(8.059)		(0.0287)	(0.0314)
Proportion of external strong ties, R^{ex}		32.13***	48.76***		-0.309***	-0.353***
		(6.104)	(8.498)		(0.0672)	(0.0770)
Geographical diversity, D		-45.22***	-44.53***		0.356*	0.394**
		(4.961)	(5.156)		(0.182)	(0.170)
$S^{\text{in}} \times R^{\text{in}}$			1.350**			-0.00464
			(0.564)			(0.00359)
$S^{\text{ex}} \times R^{\text{ex}}$			29.08***			-0.0547***
			(2.707)			(0.0157)
$S^{\text{in}} \times D$			10.68***			0.00453
			(3.179)			(0.0161)
$S^{\text{ex}} \times D$			-2.954**			0.0596***
			(1.311)			(0.00760)
Control variables						
Size, Q	427.7***	489.6***	511.0***	2.441***	2.354***	2.287***
	(28.89)	(26.98)	(28.91)	(0.181)	(0.221)	(0.225)
Centralisation, H	-27.94***	-22.75***	-26.62***	-0.399***	-0.319***	-0.244***
	(6.164)	(4.635)	(5.439)	(0.0433)	(0.0415)	(0.0421)
Betweenness centrality, BC	138.3	137.2	132.3	-0.925***	-0.877***	-0.865***
	(128.1)	(130.3)	(131.1)	(0.256)	(0.261)	(0.263)
Closeness centrality, CC	-27.75***	-44.74***	-44.09***	0.606***	0.719**	0.689**
	(4.527)	(4.116)	(4.439)	(0.227)	(0.284)	(0.270)
Focal year, t	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed
Constant	1200.0***	1209.6***	1218.8***	5.206***	5.293***	5.248***
	(170.3)	(169.5)	(168.8)	(0.708)	(0.714)	(0.717)
Observations	10897	10897	10897	10897	10897	10897
AIC	161845.6	161694.9	161595.6	63008.4	62901.1	62895.9
BIC	162137.5	162023.2	161953.1	63300.2	63229.4	63253.4
Log lik.	-80882.8	-80802.4	-80748.8	-31464.2	-31405.6	-31399.0

Robust standard errors in parentheses.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.6: Summary of regression results for hypothesis testing.

	Scientific impact, Γ		Scientific innovation, Δ	
	Hypothesis	Regression	Hypothesis	Regression
Internal brokerage, S^{in}	1(a): +	1(a): +	1(b): +	1(b): +
External brokerage, S^{ex}	1(c): −	1(c): +	1(d): −	1(d): −
Proportion of internal strong ties, R^{in}	2(a): −	2(a): NS	2(b): −	2(b): +
Proportion of external strong ties, R^{ex}	2(c): +	2(c): +	2(d): +	2(d): −
Geographical diversity, D	3(a): +	3(a): −	3(b): +	3(b): +
$S^{\text{in}} \times R^{\text{in}}$	4(a): Yes	4(a): +	4(b): Yes	4(b): NS
$S^{\text{ex}} \times R^{\text{ex}}$	4(c): Yes	4(c): +	4(d): Yes	4(d): −
$S^{\text{in}} \times D$	5(a): Yes	5(a): +	5(b): Yes	5(b): NS
$S^{\text{ex}} \times D$	5(c): Yes	5(c): +	5(d): Yes	5(d): +

+: statistically significant positive relationship; −: statistically significant negative relationship; NS: statistically not significant. Note that the hypotheses concerned with interaction effects do not specify the sign. That is why 'Yes' is used instead of a specific sign. Green (red) colour means that the hypothesis and the corresponding regression result are in agreement (disagreement). Grey colour means that regression results are statistically not significant. Regression results for hypotheses 1(a-d), 2(a-d) and 3(a-d) are based on Models 2 and 4 in Table 4.5. Regression results for hypotheses 4(a-d) and 5(a-d) are based on Models 3 and 6 in Table 4.5.

4.4.2 Regression results

Table 4.5 summarises the regression estimates from three-level hierarchical random-intercept models. Note that all the independent and control variables are standardised. Each coefficient can thus be interpreted as the expected change of the mean of the dependent variable as the independent variable varies by one standard deviation while holding all other variables constant. In Table 4.5, Models 1-3 refer to scientific impact, and Models 4-6 refer to scientific innovation. I also used information criteria (AIC and BIC) to compare Models 3 and 6 to the corresponding models that do not have the heteroskedastic residual errors. Results are in favour of the models with the heteroskedastic residual errors (AIC: 161595.6 (with residuals term) < 165367.3 (without residuals term) and BIC: 161953.1 (with residuals term) < 165608.1 (without residuals term) for Model 3; AIC: 62895.9 (with residuals term) < 70234.4 (without residuals term) and BIC: 63253.42 (with residuals term) < 70475.18 (without residuals term) for Model 6). Table 4.6 summarises the comparisons between hypotheses and empirical regression results.

Scientific impact

Model 1 I start with the baseline model, Model 1, which includes only control variables. First, as expected, the baseline model indicates that size, measured by the logarithm of the total number of internal scientists and external collaborators of a city, is positively and significantly associated with scientific impact ($\beta_1 = 427.7, p < 0.01$). Second, centralisation is negatively and significantly associated with scientific impact ($\beta_2 = -27.94, p < 0.01$). Third, the association between betweenness centrality and scientific impact is not statistically significant ($\beta_3 = 138.3, p > 0.1$). Fourth, closeness centrality is negatively and significantly associated with scientific impact ($\beta_4 = -27.75, p < 0.01$).

Model 2 To test Hypotheses 1(a), 1(c), 2(a), 2(c), and 3(a), I regress a city's scientific impact on both control and independent variables, that is, brokerage, proportion of strong ties and geographical diversity. First, both internal ($\beta_5 = 13.85, p < 0.01$) and external ($\beta_6 = 21.55, p < 0.01$) brokerage of a city are positively and significantly associated with its scientific impact. Thus, Hypothesis 1(a) is supported, and Hypothesis 1(c) is not supported. This finding suggests that the more a city brokers between external collaborators, the higher the scientific impact. This is despite the fact that brokerage may induce higher coordination and communication costs due to the geographical distance separating the external collaborators. Second, the association between the proportion of internal strong ties and scientific impact is not significant ($\beta_7 = -5.874, p > 0.1$). Hence, Hypothesis 2(a) is not supported. Moreover, the relationship between the proportion of external strong ties and scientific impact is positive and significant ($\beta_8 = 32.13, p < 0.01$), and Hypothesis 2(c) is supported. Thus, while a city gains from brokering between externals, the costs associated with lack of brokerage can be offset by the benefits arising from strongly connected collaborators, among whom communication and joint work are likely to become smoother and less costly. Third, the association between geographical diversity ($\beta_9 = -45.22, p < 0.01$) and scientific impact is negative and significant. Therefore, Hypothesis 3(a) is not supported. This indicates that by controlling for other variables, higher geographical diversity of a city's external collaborators is associated with lower scientific impact.

Model 3 Building on Model 2, in Model 3 I further add interaction terms of brokerage and proportion of strong ties, and brokerage and geographical diversity.

Model 3 thus includes all the control variables, the independent variables, and two types of interaction terms. I use Model 3 to test Hypotheses 4(a), 4(c), 5(a), and 5(c). On the one hand, in Model 3, the interaction ($\beta_{10} = 1.350, p < 0.01$) between internal brokerage and proportion of internal strong ties is positive and significant, while that ($\beta_{11} = 29.08, p < 0.01$) of external brokerage and proportion of external strong ties is also positive and significant. Hence, Hypotheses 4(a) and 4(c) are supported. On the other hand, in Model 3, the interaction ($\beta_{12} = 10.68, p < 0.01$) between internal brokerage and geographical diversity is positive and significant, and that ($\beta_{13} = -2.954, p < 0.05$) of external brokerage and geographical diversity is negative and significant. Hence, Hypotheses 5(a) and 5(c) are supported.

Scientific innovation

Model 4 As in Model 1, I begin with the baseline model, Model 4, including solely control variables. First, as expected, the baseline model suggests that the association between size and scientific innovation is positive and significant ($\beta_1 = 2.441, p < 0.01$). Second, the relationship between centralisation and scientific innovation is negative and significant ($\beta_2 = -0.399, p < 0.01$). Third, the relationship between betweenness centrality and scientific innovation is negative and significant ($\beta_3 = -0.925, p < 0.01$). Fourth, closeness centrality is positively and significantly related with scientific innovation ($\beta_4 = 0.606, p < 0.01$).

Model 5 To test Hypotheses 1(b), 1(d), 2(b), 2(d), and 3(b), similar to Model 2, after adding control variables as in Model 1, in Model 5 I regress a city's scientific innovation on the independent variables including brokerage, the proportion of strong ties and geographical diversity. First, internal brokerage ($\beta_5 = 0.145, p < 0.01$) is positively and significantly associated with scientific innovation, thus supporting Hypothesis 1(b). In addition, external brokerage ($\beta_6 = -0.175, p < 0.01$) is negatively and significantly related to scientific innovation, which is as expected. Thus, Hypothesis 1(d) is supported. Second, the association between the proportion of internal ($\beta_7 = 0.0706, p < 0.05$) strong ties and scientific innovation is positive and significant, and the relationship between the proportion of external ($\beta_8 = -0.309, p < 0.01$) strong ties and scientific innovation is negative and significant. Hence both Hypotheses 2(b) and 2(d) are rejected. Results suggest that a city can benefit from higher scientific innovation if its proportion of internal strong ties is higher while the proportion of external strong ties is lower. Third, as expected, geographical diversity ($\beta_9 = 0.356, p < 0.1$) is positively and significantly associated with scientific innovation. Hence, Hypothesis 3(b) is supported.

Model 6 Based on Model 5, in Model 6 I include interaction terms of brokerage and proportion of strong ties, and brokerage and geographical diversity. Model 6 thus includes all the control variables, independent variables, and two types of interaction terms. I use Model 6 to test Hypotheses 4(b), 4(d), 5(b), and 5(d). On the one hand, in Model 6, the interaction ($\beta_{10} = -0.00464, p > 0.1$) between internal brokerage and proportion of internal strong ties is not significant, while that ($\beta_{11} = -0.0547, p < 0.01$) of external brokerage and proportion of external strong ties is negative and insignificant. Hence, Hypothesis 4(b) is rejected, but 4(d) is supported. On the other hand, in Model 6, the interaction ($\beta_{12} = 0.00453, p > 0.1$) between internal brokerage and geographical diversity is not significant, and that ($\beta_{13} = 0.0596, p < 0.01$) of external brokerage and geographical diversity is positive and significant. Therefore, Hypothesis 5(b) is rejected, while Hypothesis 5(d) is supported.

4.4.3 Visualisation of interaction effects

To make it easier to interpret the results concerned with interactions, I visualise the interaction effects by plotting the heatmaps of predicted margins using Models 3 and 6. Results for interactions between brokerage and proportion of strong ties are shown in Figure 4.3. Results for interactions between brokerage and geographical diversity are shown in Figure 4.4. In each panel in Figures 4.3 and 4.4, scientific impact or innovation is predicted as a function of two focal variables while keeping other variables at their means. The interaction terms in my models are responsible for the curvature of the contour lines in the panels in Figures 4.3 and 4.4. Without interactions, the contour lines between different levels of colours would be straight (see Figures 4.3b and 4.4b). The curvature demonstrates that the relationship between the (internal or external) brokerage and scientific impact or innovation differs across levels of proportion of (internal or external) strong ties or geographical diversity, and vice versa.

4.5 Discussion and conclusion

In this study, I mainly found that (i) the same independent covariate (e.g., geographical diversity) may have different relationships with different performance-related outcomes of a city, thus highlighting the importance of considering scientific performance from different angles; (ii) the internal and external measures of

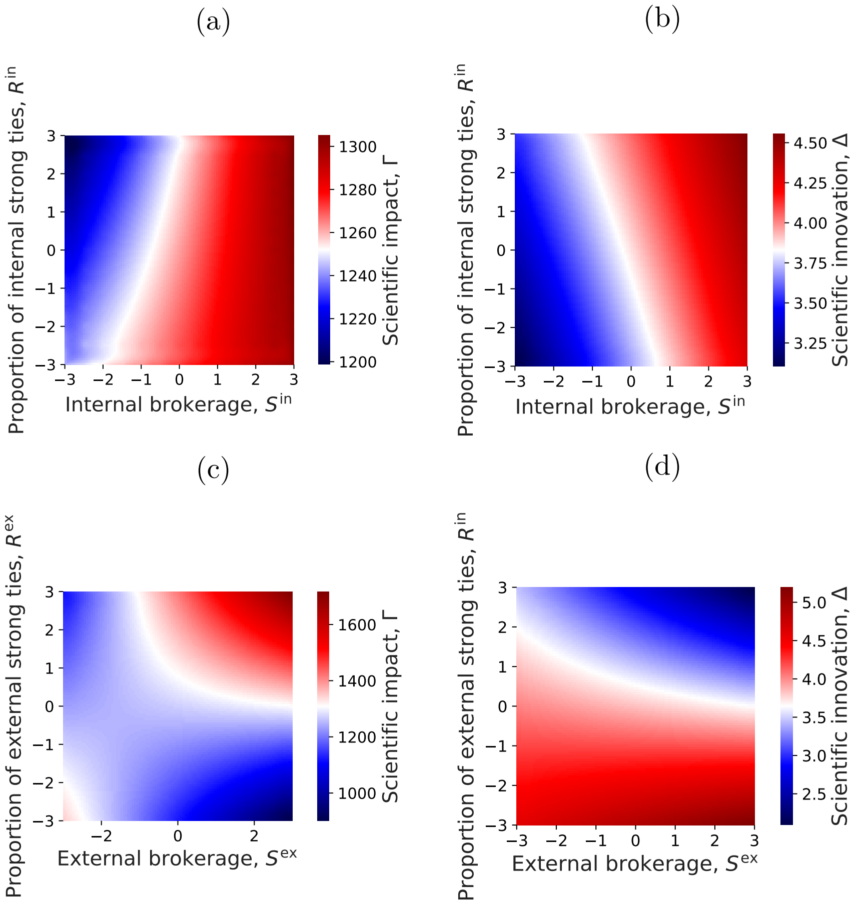


Figure 4.3: **Visualisation of interaction effects between brokerage and proportion of strong ties in Models 3 and 6 in Table 4.5.** The figure plots the scientific impact (Panels (a) and (c)) and innovation (Panels (b) and (d)) of a city predicted as a function of internal or external brokerage (standardised), and proportion of strong ties (standardised).

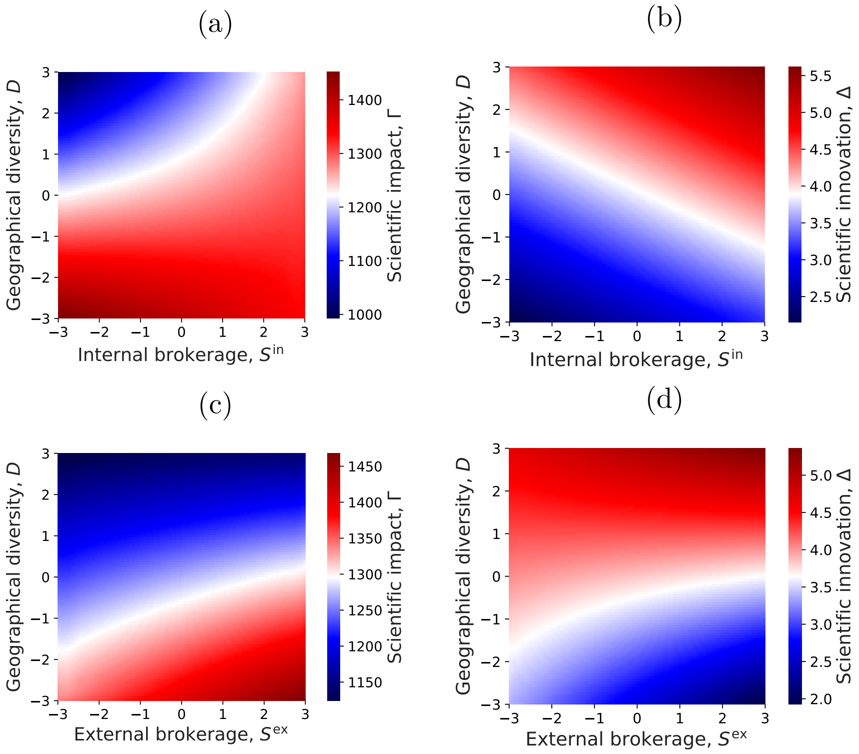


Figure 4.4: **Visualisation of interaction effects between brokerage and geographical diversity in Models 3 and 6 in Table 4.5.** The figure plots the scientific impact (Panels (a) and (c)) and innovation (Panels (b) and (d)) of a city predicted as a function of internal or external brokerage (standardised), and geographical diversity (standardised).

social capital (e.g., brokerage) may have distinct associations with the same performance-related outcome, thus suggesting that my proposed finer-grained measures can capture different perspectives of network collaboration patterns; and (iii) estimates of interaction effects reveal that these measures do not play their roles independently. Specifically, I found that brokerage can be moderated by the proportion of strong ties and geographical diversity.

4.5.1 Implications for research

Prior studies (Guan et al., 2015) concerned with the relationship between scientific collaboration networks and the scientific performance of cities focused mainly on the inter-city collaboration network where nodes are the cities and links represent the collaboration between cities. This type of method ignores the actual interaction patterns among all scientists somehow related to cities. For example, two cities with equivalent topological structures in the inter-city collaboration network may have dramatically different collaboration patterns among their internal scientists. This limitation has been addressed in this chapter by proposing that a city can be explicitly expressed as a function of the collaboration network of its internal scientists and external collaborators through several network measures to quantify its social capital. By assembling a large-scale longitudinal and global data set from Medline, Author-ity, MapAffi, and SCImago, I studied how network characteristics of the collaboration network of internal scientists and external collaborators of a city are associated with the city's scientific impact and innovation. The approach I proposed here to constructing collaboration networks can be adopted in future work concerned with new network measures and their relationships with the scientific performance of cities.

In this work, I focused on three sets of measures: (i) the brokerage potential, a new measure that builds on, and extends, Burt's theory of structural holes (Burt, 2009); (ii) the proportion of strong ties, a measure inspired by Granovetter's theory of the strength of weak ties (Granovetter, 1973); and (iii) geographical diversity, increasingly advocated as the new orthodoxy in city planning (Fainstein, 2005). Notice that my methodology allows me to define finer-grained internal and external brokerage measures and the proportion of strong ties by drawing attention separately to internal scientists and external collaborators of a city. The main benefit of introducing internal and external measures is that one can potentially offer complementary perspectives on structural sources of social capital. This will help other researchers to consider finer-grained sources of social capital in future research and obtain a deeper understanding of their (potentially distinct) relationships with performance-based outcomes.

Findings based on the regression analysis contribute to the ongoing debate on social capital and its foundations. First, results suggest that the same social capital measure may have different associations with different performance-related outcomes of a city. Specifically, I notice that external brokerage is significantly and positively associated with scientific impact but is significantly and negatively related to scientific innovation. This indicates that, while holding other variables unchanged, an increase of a city's external brokerage is expected to be associated with a higher scientific impact but with lower scientific innovation. The proportion of external strong ties is positively associated with scientific impact and negatively related to scientific innovation. In addition, geographical diversity is negatively associated with scientific impact, but that is positively associated with scientific innovation. Second, the internal and external measures may have distinct associations with the same performance-related outcome, suggesting that my proposed finer-grained measures can capture different perspectives of network collaboration patterns. Specifically, internal and external brokerage measures show distinct relationships with scientific innovation. The proportions of internal and external strong ties also have different associations with scientific innovation. This is in agreement with the argument in the social sciences that simply considering only the network structure while ignoring the properties of the actors may not provide a comprehensive perspective on the structural foundations of social capital (Aral and Van Alstyne, 2011; Fleming et al., 2007; Schilling and Fang, 2014; Shipilov and Li, 2008; Ter Wal et al., 2016; Uzzi, 1996). Indeed my study suggests that non-topological properties of the interacting nodes need to be taken into account. For example, in scientific collaboration networks among scientists, the place of residence (i.e., the city) of a scientist can be seen as the non-topological property of the node.

Furthermore, my study can be seen as the first attempt to apply the concept of interconnected geo-social network proposed in Ref. (Hristova et al., 2016) to spatial networks in the community of research policy. This consists of two interconnected network layers between people (social layer) and places (geographical layer). On the one hand, the scientific collaboration network of scientists corresponds to the social layer. On the other hand, the inter-city collaboration network I constructed to compute a city's betweenness centrality and closeness centrality can be regarded as the geographical layer. For the geographical layer, instead of considering the collaboration between cities, other types of interactions could be addressed in future work. For example, by drawing on scientists' mobility data (Edler et al., 2011; Scellato et al., 2015; Verginer and Riccaboni, 2020a,b), the inter-city collaboration network can be constructed in which the cities are the nodes and links between cities reflect the mobility flows between them. This will open up new avenues for studying the association between scientific performance and scientists' inter-city mobility and collaboration.

4.5.2 Implications for practice

Apart from the theoretical implications of my findings, my study can also provide practical implications for scientists, policymakers, planning agencies and governments.

On the one hand, I can assess the longitudinal rankings of cities over time in terms of the different perspectives on social capital I proposed, which can allow policymakers to carry out a comparative analysis of cities in the world. The rankings can be used as partial guidance to inform scientists' decisions about when and where to move and can be used retrospectively to assess if research policies have produced the desired effect in promoting a city as a prime research location for policymakers. Moreover, the rankings of cities can allow scientists and policymakers to study and predict the emergence and disappearance of scientific hubs (cities or regions).

On the other hand, based on scientific collaboration networks, my analysis identifies a number of factors associated with cities' scientific impact and innovation. This can provide policymakers with insights on how to maintain or improve the scientific performance of cities. First, policies need to focus not only on internal scientists but also on external collaborators of cities. As I have shown, both types of scientists and their interactions play essential roles in contributing to the scientific performance of cities. Second, policymakers should pay attention to the interplay among brokerage, strong ties, and geographical diversity. Proper strategies should be made according to the comprehensive judgement of multiple proposed indicators of the social capital of cities. Third, scientific impact and scientific innovation are different perspectives of scientific performance. The same change in collaboration patterns may lead to different or even opposite influences on these two kinds of outcomes. Policymakers should make their objectives clear in terms of developing the scientific performance of cities, e.g., by prioritising impact or encouraging innovation.

4.5.3 Limitations

In what follows I will describe limitations concerned with my study, and then outline how these might be overcome in future work. First, the main data set to obtain scientific publications in my study is Medline, a bibliographic platform focusing on life sciences and biomedical information. In my future study, I should also consider other bibliographic platforms (e.g., Web of Science and Microsoft Academic Graph (Sinha et al., 2015)) to include a range of disciplines beyond those in Medline. For example, Microsoft Academic Graph contains 19 different scientific fields, offering opportunities for us to uncover universal regularities across fields as well as domain-specific patterns. Second, by focusing on bibliometric data, this study is limited to scientific collaboration concerned with publications without fully capturing the collaboration between scientists beyond the publications, e.g., grants, patents, or other research activities, which should be incorporated in future work. Third, scientific impact and innovation are considered as measures of scientific performance in my study. Other important measures that could capture scientific performance (e.g., patents and funding) should also be included in the analysis in future work. Fourth, demographic characteristics of scientists (e.g., gender, age and ethnicity) should also be incorporated into future studies by drawing on publicly available large-scale demographic data sets, such as Ref. (Ke et al., 2021). Incorporating characteristics of scientists in future work can offer more insights about the relationships between the proposed social capital measures and scientific performance especially in studies where the unit of analysis is a scientist. Fifth, while this study focuses on the city level, it would also be interesting to replicate the study on the institution level, compare whether results are consistent, and understand whether the institution level can offer finer-grained insights in future work. The data used in the study does not allow me to focus on the institution level because the granularity of geographical information about scientists is only at the city level. Sixth, my study focuses on examining the relationships between the proposed independent measures and dependent measures. It thus does not provide causal interpretations of the relationship between them, and I have paid careful attention to avoiding using any causal terms throughout the study.

4.6 Contribution to the literature

Previous studies have used the inter-city scientific collaboration network to understand the relationship between the social capital of a city and its scientific performance (Guan et al., 2015). Here in the inter-city collaboration network, nodes are cities, and two cities are connected if scientists from two cities have co-authored at least one paper. However, the inter-city collaboration network aggregated from the individual scientist level lacks the actual collaboration patterns of scientists within and across cities. In this chapter, my proposed geo-social network approach (see Figure 4.1) that constructs collaboration networks of internal scientists and external collaborators associated with a city can address the above limitation and offers a methodological contribution to the studies concerned with the interplay of networks, geography, and scientific performance.

To exemplify the applicability of the proposed approach, I applied it to large-scale bibliometric data sets and quantified finer-grained measures to uncover the social capital of a city extracted from the collaboration networks of its internal scientists and external collaborators. I further studied how the finer-grained measures of social capital of a city are associated with its scientific performance from two distinct perspectives: impact and innovation. The relationship between finer-grained measures of social capital and scientific performance has been examined. Although I focused on cities as the unit of analysis in this chapter, the proposed methodology can also contribute to other studies concerned with different levels of analysis, including departments, institutions, and whole countries.

Table 4.1: Calculation of the measures for London in Figure 4.1

Measure	Symbol	Result
Set of internal scientists and external collaborators	$\mathcal{N}_{\mathrm{S}}$	$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_{10}\}$
Set of internal scientists	$\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}$	$\{u_1, u_2, u_3, u_4, u_5\}$
Set of external collaborators	$\mathcal{N}_{\mathrm{S}}^{\mathrm{ex}}$	$\{u_6, u_7, u_{10}\}$
Size	$\lvert\mathcal{N}_{\mathrm{S}}\rvert$	8
Internal size	$\lvert\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}\rvert$	5
External size	$\lvert\mathcal{N}_{\mathrm{S}}^{\mathrm{ex}}\rvert$	3
Centralisation	H	In the toy example in Figure 4.1, u^* is u_3. $H(\text{London}) = [6 \times 8 - (3+2+6+2+1+3+3+2)]/[(8-1)(8-2)] = 0.619$.
Internal brokerage	S^{in}	$[5 - (2 \times 5)/5]/5 = 0.6$
External brokerage	S^{ex}	$[3 - (2 \times 2)/3]/3 = 0.556$
Set of resident cities of external collaborators	$\mathcal{N}_{\mathrm{L}}$	$\{\text{Boston}, \text{Lucca}\}$
Geographical diversity	D	In the toy example in Figure 4.1, I assume the weights of all links are 1. But in real calculations, I take the weights of links as defined above. In this case, $\mathcal{N}_{\mathrm{L}}(\text{London}) = \{\text{Boston}, \text{Lucca}\}$. $P(\text{London}, \text{Boston}) = 3/4$. $P(\text{London}, \text{Lucca}) = 1/4$. $D(\text{London}) = 1 - (3/4)^2 - (1/4)^2 = 0.375$.

Table 4.2: **Summary of main notations.**

Notation	**Description**
u, v	Scientists.
i, j	Cities.
t	The focal year.
$\mathcal{T}_t$	The focal window $\mathcal{T}_t = [y, y+\tau)$ where t is the focal year and τ is the window size (years).
$w_{\mathrm{S}}(u, v)$	The weight of the collaboration link of scientists u and v in $\mathcal{T}_t$.
$w_{\mathrm{L}}(i, j)$	The aggregated value of weights of collaboration links of scientists resident in cities i and j, respectively. It is also the weight of the collaboration link of cities i and j in the inter-city collaboration network in $\mathcal{T}_t$.
$L(u)$	The resident city of scientist u in $\mathcal{T}_t$.
$\Gamma(i)$	Scientific impact of city i in $\mathcal{T}_t$.
$\Delta(i)$	Scientific innovation of city i in $\mathcal{T}_t$.
$\mathcal{G}(i)$	The collaboration network of internal scientists and external collaborators of city i in $\mathcal{T}_t$.
$\mathcal{G}^{\mathrm{in}}(i)$	The collaboration network of internal scientists of city i in $\mathcal{T}_t$.
$\mathcal{G}^{\mathrm{ex}}(i)$	The collaboration network of external collaborators of city i in $\mathcal{T}_t$.
$\mathcal{N}_{\mathrm{S}}(i)$	The set of internal scientists and external collaborators of city i in $\mathcal{T}_t$.
$\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)$	The set of internal scientists of city i in $\mathcal{T}_t$.
$\mathcal{N}_{\mathrm{S}}^{\mathrm{ex}}(i)$	The set of external collaborators of city i in $\mathcal{T}_t$.
$\lvert\mathcal{N}_{\mathrm{S}}(i)\rvert$	The total number of internal scientists and external collaborators of city i in $\mathcal{T}_t$.
$\lvert\mathcal{N}_{\mathrm{S}}^{\mathrm{in}}(i)\rvert$	The number of internal scientists of city i in $\mathcal{T}_t$.
$\lvert\mathcal{N}_{\mathrm{S}}^{\mathrm{ex}}(i)\rvert$	The number of external scientists of city i in $\mathcal{T}_t$.
$\mathcal{N}_{\mathrm{L}}(i)$	The set of resident cities of external collaborators of city i in $\mathcal{T}_t$. It is also the set of neighbours of city i in the inter-city collaboration network in $\mathcal{T}_t$.
$BC(i)$	Betweenness centrality of city i in the inter-city collaboration network $\mathcal{T}_t$.
$CC(i)$	Closeness centrality of city i in the inter-city collaboration network $\mathcal{T}_t$.
$H(i)$	Centralisation of city i in in $\mathcal{T}_t$.
$S^{\mathrm{in}}(i)$	Internal brokerage of city i in $\mathcal{T}_t$.
$S^{\mathrm{ex}}(i)$	External brokerage of city i in $\mathcal{T}_t$.
$R^{\mathrm{in}}(i)$	Proportion of internal strong ties of city i in $\mathcal{T}_t$.
$R^{\mathrm{ex}}(i)$	Proportion of external strong ties of city i in $\mathcal{T}_t$.
$D(i)$	Geographical diversity of city i in $\mathcal{T}_t$.

Part II: Deep learning

Chapter 5

Background

Part II of my thesis will be concerned with two projects on graph-based deep learning and classification tasks. The idea of graph-based deep learning methods is to learn low-dimensional representations of nodes from original high-dimensional data features by incorporating both node features and graph structure that describes relational information between nodes. This area of investigation is somewhat conceptually related to my projects on social capital if we regard a measure for extracting sources of social capital as a function to map a node's ego-centred network structure and/or network metadata (e.g., node attributes) to a numeric value, i.e., a one-dimensional representation of a node. For example, my proposed new brokerage measure in Chapter 3 quantifies an actor's brokerage opportunity (a numeric value) as a function of network patterns between nodes and their attributes in this actor's ego-centred network. In what follows, I shall start with the background related to graph-based deep learning.

Deep learning encompasses a broad class of machine learning methods that use multiple layers of nonlinear processing units in order to learn multi-level representations for detection or classification tasks (Bronstein et al., 2017; Deng and Yu, 2014; Goodfellow et al., 2016; LeCun et al., 2015; Schmidhuber, 2015). The main realisations of deep multi-layer architectures are the so-called Deep Neural Networks (DNNs), which correspond to Artificial Neural Networks (ANNs) with multiple layers between input and output layers. DNNs have been shown to perform successfully in processing a variety of signals with an underlying Euclidean or grid-like structure, such as speech, images, and videos. Signals with an underlying Euclidean structure usually come in the form of multiple arrays (LeCun et al., 2015) and are known for their statistical properties such as locality, stationarity, and hierarchical compositionality from local statistics (Field, 1989; Simoncelli and Olshausen, 2001). For instance, an image can be seen as a function on Euclidean space (the 2D plane) sampled from a grid. In this setting, the locality is a consequence of local connections, stationarity results from shift-invariance, and compositionality stems from the intrinsic multi-resolution structure of many images (Bronstein et al., 2017). It has been suggested that such statistical properties can be exploited by convolutional architectures via DNNs, namely (deep) Convolutional Neural Networks (CNNs) (Bruna and Mallat, 2013; LeCun et al., 1990, 1998) which are based on four main ideas: local connections, shared weights, pooling, and multiple layers (LeCun et al., 2015). The role of the convolutional layer in a typical CNN architecture is to detect local features from the previous layer that are shared across the image domain, thus largely reducing the parameters compared with traditional fully connected feed-forward ANNs.

Although deep learning models, and in particular CNNs, have achieved highly improved performance on data characterised by an underlying Euclidean structure, many real-world data sets do not have a natural and direct connection with a Euclidean space. Recently there has been interest in extending deep learning techniques to non-Euclidean domains, such as graphs and manifolds (Bronstein et al., 2017). An archetypal example is social networks, represented as graphs with users as nodes and edges representing social ties between them. In biology, gene regulatory networks represent relationships between genes encoding proteins that can up- or down-regulate the expression of other genes. Here I review the development of extending neural networks on graphs.

5.1 Early developments of neural networks on graphs

The first attempt to generalise neural networks on graphs can be traced back to Ref. (Gori et al., 2005), who proposed a scheme combining Recurrent Neural Networks (RNNs) and random walk models. Their method requires the repeated application of contraction maps as propagation functions until the node representations reach a stable fixed point. This method, however, did not attract much attention when it was proposed. With the current surge of interest in deep learning, this work has been reappraised in a new and modern form: Ref. (Li et al., 2016) introduced modern techniques for RNN training based on the original GNN framework, whereas Ref. (Duvenaud et al., 2015) proposed a convolution-like propagation rule on graphs and methods for graph-level classification. Non-spectral methods have also been successfully proposed. For example, Ref. (Atwood and

Towsley, 2016) shows how diffusion-based representations can be learned from graph-structured data and used as the basis for node classification by introducing a diffusion-convolution operation. Ref. (Niepert et al., 2016) converts graphs locally into sequences fed into a conventional 1D CNN, which needs the definition of a node ordering in a pre-processing step.

The first formulation of CNN on graphs (GCNNs) was proposed by Ref. (Bruna et al., 2014). These researchers applied the definition of convolutions to the spectral domain of the graph Laplacian, which is presented in Section 5.2.

5.2 Spectral graph convolutions

I now briefly present the key insights introduced by Ref. (Bruna et al., 2014) to extend CNNs to the non-Euclidean domain. For an extensive recent review, the reader should refer to Ref. (Bronstein et al., 2017).

In a data set, each sample is described by a C^0-dimensional feature vector, which is conveniently arranged into the feature matrix $X \in R^{N \times C^0}$. Each sample is also associated with the node of a given graph $\mathcal{G}$ with N nodes, with edges representing additional relational (symmetric) information. This undirected graph is described by the adjacency matrix $A \in R^{N \times N}$. The ground truth assignment of each node to one of F classes is encoded into a 0-1 membership matrix $Y \in R^{N \times F}$.

The main hurdle is the definition of a convolution operation on a graph between a filter g_θ and the node features X. In signal processing, a filter refers to a process to remove some undesirable components or features from a signal. This can be achieved by expressing g_θ onto a basis encoding information about the graph with nodes $v_i \in V$ and edges $(v_i, v_j) \in E$, e.g., the adjacency matrix A or the normalised Laplacian $L = I_N - D^{-1/2}AD^{-1/2}$, where I_N is the identity matrix and $D = \mathrm{diag}(A\mathbf{1})$. This real symmetric matrix has an eigendecomposition $L = U\Lambda U^T$, where U is the matrix of column eigenvectors with associated eigenvalues collected in the diagonal matrix Λ. The spectral convolutions on graphs are considered as a multiplication between a signal $x \in R^N$ (a scalar for every node) and a filter $g_w = \mathrm{diag}(\theta)$ parameterised by $\theta \in R^N$ in the Fourier domain:

$$g_\theta \star X = U g_\theta(\Lambda) U^T X, \tag{5.1}$$

where $U^T X$ represents the graph Fourier transforms of signals X on nodes. $g_\theta(\Lambda)$ can be understood as a function of eigenvalues of L, filtering $U^T X$ in the frequency domain. Finally, the signal is projected back onto the nodes by multiplying U on the left.

While being theoretically salient, this method is unfortunately impractical due to its computational complexity. Specifically, evaluating Equation (5.1) is computationally inefficient because multiplication with the eigenvector matrix U is $O(N^2)$. In addition, computing the eigendecomposition of L is also expensive for large graphs. To address this problem, authors of Ref. (Hammond et al., 2011) suggest that $g_\theta(\Lambda)$ can be well-approximated by a truncated expansion according to Chebyshev polynomials $T_k(x)$ up to K^{th} order:

$$g_{\theta'}(\Lambda) \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{\Lambda}), \tag{5.2}$$

where $\tilde{\Lambda} = \frac{2}{\lambda_{\max}}\Lambda - I_N$. $\lambda_{\max}$ represents the largest eigenvalue of L. $\theta' \in R^K$ is a vector of Chebyshev coefficients. The Chebyshev polynomials are recursively defined as $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ where $T_0(x) = 1$ and $T_1(x) = x$. This can be integrated into the definition of a convolution of a signal x with a filter $g_{\theta'}$:

$$g_{\theta'} \star x \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{L})x, \tag{5.3}$$

with $\tilde{L} = \frac{2}{\lambda_{\max}}L - I_N$. It can be noticed that this expression is now K-localised because it is a K^{th}, i.e., it depends only on nodes that are at maximum K steps away from the central node. The computational complexity of considering Equation (5.3) is $O(|E|)$, i.e., linear in the number of edges. Ref. (Defferrard et al., 2016) leveraged this K-localised convolution to define a convolutional neural network on graphs. In Ref. (Kipf and Welling, 2017), a GCN architecture was proposed via a first-order approximation of localised spectral filters on graphs, i.e., $K = 1$ in Equation (5.3). In this linear formulation of a GCN, $\lambda_{\max}$ is approximated to be 2. Under these approximations, Equation (5.3) is further simplified into:

$$g_{\theta'} \star x \approx \theta'_0 x + \theta'_1 (L - I_N) x = \theta'_0 x - \theta'_1 D^{-\frac{1}{2}} A D^{-\frac{1}{2}} x \tag{5.4}$$

with two free parameters θ'_0 and θ'_1. The authors of Ref. (Kipf and Welling, 2017) further constrain the number of parameters in Equation (5.4) and obtain the following expression:

$$g_\theta \star x \approx \theta \left(I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \right) x \tag{5.5}$$

with a single parameter $\theta = \theta'_0 = -\theta'_1$. It has been noticed that $I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ now has eigenvalues in the range $[0, 2]$. Repeated applications of this operator might lead to numerical instabilities and exploding/vanishing gradients in deep neural networks. To overcome this issue, the following renormalisation trick is introduced: $I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \rightarrow \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$, where $\tilde{A} = A + I_N$ and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. This definition can be generalised to a signal $X \in R^{N \times C^0}$ and F filters as follows:

$$Z = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X \Theta \tag{5.6}$$

where $\Theta \in R^{C^0 \times F}$ is a matrix of filter parameters and $Z \in R^{N \times F}$ is the convolved signal matrix. This operation has complexity $O(|E|FC)$ because $\tilde{A}X$ can be efficiently implemented as a product of a sparse matrix with a dense matrix.

The authors of Ref. (Kipf and Welling, 2017) considered the task of semi-supervised transductive node classification, where labels are only available for a small number of nodes. Starting with a feature matrix X and a network adjacency matrix A, they encoded the graph structure directly using a neural network model $f(X, A)$, and trained on a supervised target loss function $\mathcal{L}$ computed over the subset of nodes with known labels. Their proposed GCN was shown to achieve improved accuracy in classification tasks on several benchmark citation networks and a knowledge graph data set. The architecture and propagation rules of this method based on Equation (5.6) are detailed in Section 5.3.

5.3 Graph Convolutional Networks

5.3.1 Layer-wise propagation rule and multi-layer architecture

Given the matrix X with sample features and the (undirected) adjacency matrix A of the graph $\mathcal{G}$ encoding relational information between the samples, the propagation rule between layers ℓ and $\ell+1$ (of size C^{ℓ} and $C^{\ell+1}$, respectively) is given by:

$$H^{\ell+1} = \sigma^{\ell}\left(\widehat{A}H^{\ell}W^{\ell}\right), \tag{5.7}$$

where $H^{\ell} \in R^{N \times C^{\ell}}$ and $H^{\ell+1} \in R^{N \times C^{\ell+1}}$ are matrices of activation in the ℓ^{th} and $(\ell+1)^{th}$ layers, respectively; $\sigma^{\ell}(\cdot)$ is the threshold activation function for layer ℓ; and the weights connecting layers ℓ and $\ell+1$ are stored in the matrix $W^{\ell} \in R^{C^{\ell} \times C^{\ell+1}}$. Note that the input layer contains the feature matrix $H^{0} \equiv X$.

5.3.2 Semi-supervised node classification

In a semi-supervised learning setting, a small subset of the node ground truth labels is used in the training phase to infer the class of unlabelled nodes. This type of learning paradigm, where only a small amount of labelled data is combined with a large amount of unlabelled data during training, lies between supervised and unsupervised learning.

Node classification on a graph using a GCN can be seen as a label propagation task: given a set of seed nodes with known labels, the task is to predict which label will be assigned to the unlabelled nodes given a certain topology and node attributes.

Following (Kipf and Welling, 2017), I implement a two-layer GCN with propagation rule (Equation (5.7)) and different activation functions for each layer, i.e., a rectified linear unit for the first layer and a softmax unit for the output layer:

$$\sigma^0 : \mathrm{ReLU}(x_i) = \max(x_i, 0) \tag{5.8}$$

$$\sigma^1 : \mathrm{softmax}(x)_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}, \tag{5.9}$$

where x is a vector. The model then takes the simple form:

$$Z = f(X, A) = \mathrm{softmax}(\widehat{A}\ \mathrm{ReLU}(\widehat{A}XW^0)\ W^1), \tag{5.10}$$

where the softmax function is applied row-wise, and the ReLU is applied element-wise. Note that there is only one hidden layer with C^1 units. Hence $W^0 \in R^{C^0 \times C^1}$ maps the input with C^0 features to the hidden layer and $W^1 \in R^{C^1 \times C^2}$ maps these hidden units to the output layer with $C^2 = F$ units, corresponding to the number of classes of the ground truth.

In this semi-supervised multi-class classification, the cross-entropy error over all labelled instances is evaluated as follows:

$$\mathcal{L} = -\sum_{l \in \mathcal{Y}_L} \sum_{f=1}^{F} Y_{lf}\ \ln Z_{lf}, \tag{5.11}$$

where $\mathcal{Y}_L$ is the set of nodes that have labels. The weights of the neural network (W^0 and W^1) are trained using gradient descent to minimise the loss $\mathcal{L}$. A visual

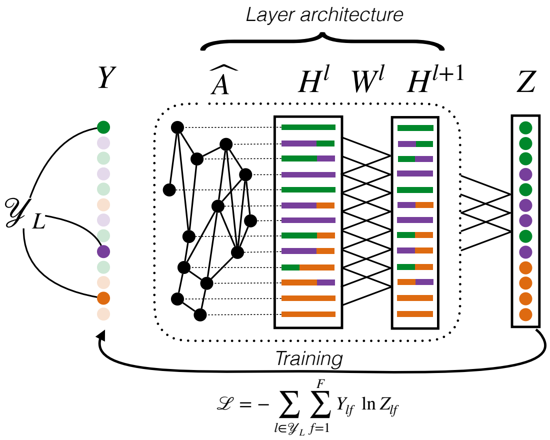


Figure 5.1: **Schematic illustration of the Graph Convolutional Network used.** The graph $\widehat{A}$ is applied to the input of each layer ℓ before it is funnelled into the input of layer $\ell + 1$. The process is repeated until the output has dimension $N \times F$ and produces a predicted class assignment. During the training phase, the predicted assignments are compared against a subset of values $\mathcal{Y}_L$ of the ground truth.

summary of the GCN architecture is shown in Figure 5.1.

Chapter 6

Quantifying the alignment of graph and features

6.1 Introduction

To address the challenge of extending deep learning techniques to graph-structured data, a new class of deep learning algorithms, broadly named GNNs, has been recently proposed (Bronstein et al., 2017; Hamilton et al., 2017; Xu et al., 2019). In this setting, each node of the graph represents a sample described by a feature vector, and I am additionally provided with relational information between the samples that can be formalised as a graph. GNNs are well suited to node (i.e., sample) classification tasks. For a recent survey of this fast-growing field, see Ref. (Wu et al., 2020).

Generalising convolutions to non-Euclidean domains is not straightforward (Defferrard et al., 2016). Recently, GCN has been proposed (Kipf and Welling, 2017) as a subclass of GNNs with convolutional properties. The GCN architecture combines the full relational information from the graph together with the node features to accomplish the multi-class classification task, using the ground truth class assignment of a small subset of nodes during the training phase. GCNs have shown improved performance for semi-supervised classification of documents (described by their text) into topic areas, outperforming methods that rely exclusively on text information without the use of any citation information, e.g., MLP (Kipf and Welling, 2017).

However, one would not expect such an improvement to be universal. In some cases, the additional information provided by the graph (i.e., the edges) might not be consistent with the similarities between the features of the nodes. In particular, in the case of citation graphs, it is not always the case that documents cite other documents that are similar in content. As I will show below with some illustrative data sets, in those cases, the conflicting information provided by the graph means that a graph-less MLP approach outperforms GCN. Here, I explore the relative importance of the graph with respect to the features for classification purposes, and propose a geometric measure based on subspace alignment to explain the relative performance of GCN against different limiting cases.

My hypothesis is that a degree of alignment among the three layers of information available (i.e., the features, the graph and the ground truth) is needed for GCN to perform well and that any degradation in the information content leads to an increased misalignment of the layers and worsened performance. I will first use randomisation schemes to show that the systematic degradation of the information contained in the graph and the features leads to a progressive worsening of GCN performance. Second, I propose a simple spectral alignment measure and show that this measure correlates with the classification performance in a number of data sets: (i) a constructive example built to illustrate my work; (ii) Cora, a well-known citation network benchmark; (iii) AMiner, a newly constructed citation network data set; and (iv) two subsets of Wikipedia: Wikipedia I, where GCN outperforms MLP, and Wikipedia II, where instead MLP outperforms GCN.

6.2 Methods

6.2.1 Randomisation strategies

To test the hypothesis that a degree of alignment across information layers is crucial for a good classification performance of GCN, I gradually randomise the node features, the node connectivity, or both. For the randomisation to give a meaningful notion of alignment, at least one ingredient needs to be kept constant. Since I focus on the alignment of graph and features, I keep the ground truth constant.

Randomisation of the graph

The edges of the graph are randomised by rewiring a percentage $p_{\widehat{A}}$ of edge stubs (i.e., “half-edges”) under the constraint that the degree distribution remains unchanged. This randomisation strategy is described in Algorithm 1 which is based on the configuration model (Newman, 2003). Once a randomised realisation of the graph is produced, the corresponding $\widehat{A}$ is computed.

```
Algorithm 1: Randomisation of the graph
  Input: A graph $G(V, E)$, where $V$ is the set of nodes and $E$ is the set of
         edges, and a randomisation percentage $0 \leq p_{\widehat{A}} \leq 100$.
  Output: A randomised graph $G_{p_{\widehat{A}}}(V, E')$
  1.Choose a random subset of edges $E_r$ from $E$ with $|E_r| = \left\lfloor |E| \times p_{\widehat{A}}/100 \right\rfloor$,
    and denote the unrandomised edges in $E$ as $E_u$.
  2. Obtain the degree sequence of nodes from $E_r$, and build a stub list $l_s$
    based on the degree sequence.
  3. Obtain a randomised stub list $l'_s$ by shuffling $l_s$, and randomised edges
    $E'_r$ by connecting the stubs in the corresponding positions of the two stub
    lists $l_s$ and $l'_s$.
  4. Compute $E_u \cup E'_r$, remove multiedges and self-loops, and obtain the
    final edge set E'.
  5. Generate randomised graph $G_{p_{\widehat{A}}}(V, E')$ from node set V and edge set
    $E'$.
```

Randomisation of the features

The features were randomised by swapping feature vectors between a percentage p_X of randomly chosen nodes following the procedure described in Algorithm 2.

```
Algorithm 2: Randomisation of the features
  Input: A feature matrix $X \in R^{N \times C^0}$, and a randomisation percentage
         $0 \leq p_X \leq 100$.
  Output: A randomised feature matrix $X_{p_X} \in R^{N \times C^0}$
  1. Choose at random $N_r$ rows from $X$, where $N_r = \lfloor N\, p_X/100 \rfloor$.
  2. Swap randomly the $N_r$ rows to obtain $X_{p_X}$.
```

A fundamental difference between the two randomisation schemes is that the graph randomisation alters its spectral properties as it gradually destroys the graph structure, whereas the randomisation of the features preserves its spectral properties in the principal component analysis (PCA) sense, i.e., the principal values are the same, but the loadings on the components are swapped. Hence the feature randomisation still alters the classification performance because the features are re-assigned to nodes that have a different environment, thereby changing the result of the convolution operation defined by the H^ℓ activation matrices (Equation (5.7)).

6.2.2 Limiting cases

To interrogate the role that the graph plays in the classification performance of a GCN, it is instructive to consider three limiting cases:

- *No graph:* $A = \mathbf{0}\mathbf{0}^T$. If I remove all the edges in the graph, the classifier becomes equivalent to an MLP, a classic feed-forward ANN. The classification is based solely on the information contained in the features, as no graph structure is present to guide the label propagation.
- *Complete graph:* $A = \mathbf{1}\mathbf{1}^T - I_N$. In this case, the mixing of features is immediate and homogeneous, corresponding to a mean field approximation of the information contained in the features.
- *No features:* $X = I_N$. In this case, the label propagation and assignment are purely based on graph topology.

An illustration of these limiting cases can be found in the top row of Table 6.2.

6.2.3 Spectral alignment measure

In order to quantify the alignment between the features, the graph, and the ground truth, I propose a measure based on the chordal distance between subspaces, as follows.

Chordal distance between two subspaces

Recent work by Ref. (Ye and Lim, 2016) has shown that the distance between two subspaces of different dimensions in $\mathbb{R}^n$ is necessarily defined in terms of their principal angles.

Let $\mathcal{A}$ and $\mathcal{B}$ be two subspaces of the ambient space $\mathbb{R}^n$ with dimensions α and β, respectively, with $\alpha \leq \beta < n$. The principal angles between $\mathcal{A}$ and $\mathcal{B}$ denoted $0 \leq \theta_1 \leq \theta_2 \leq ... \leq \theta_\alpha \leq \frac{\pi}{2}$ are defined recursively as follows (Björck and Golub, 1973; Golub and Van Loan, 2013):

$$\begin{aligned}
\theta_1 &= \min_{a_1 \in \mathcal{A}, b_1 \in \mathcal{B}} \arccos\left(\frac{|a_1^T b_1|}{\|a_1\| \|b_1\|}\right), \\
\theta_j &= \min_{\substack{a_j \in \mathcal{A}, b_j \in \mathcal{B} \\ a_j \perp a_1, ..., a_{j-1} \\ b_j \perp b_1, ..., b_{j-1}}} \arccos\left(\frac{|a_j^T b_j|}{\|a_j\| \|b_j\|}\right), \quad j = 2, ..., \alpha,
\end{aligned}$$

If the minimal principal angle is small, then the two subspaces are nearly linearly dependent, i.e., almost perfectly aligned. A numerically stable algorithm that computes the canonical correlations (i.e., the cosine of the principal angles) between subspaces is given in Algorithm 3.

Algorithm 3: Principal angles (Björck and Golub, 1973; Golub and Van Loan, 2013)

```
Input: matrices $A_{n \times \alpha}$ and $B_{n \times \beta}$ with $\alpha \leq \beta < n$.
Output: cosines of the principal angles $\theta_1 \leq \theta_2 \leq ... \leq \theta_\alpha$ between $\mathcal{R}(A)$ and $\mathcal{R}(B)$, the column spaces of $A$ and $B$.
1. Find orthonormal bases $\mathcal{Q}_A$ and $\mathcal{Q}_B$ for $A$ and $B$ using the QR decomposition: $\mathcal{Q}_A^T \mathcal{Q}_A = \mathcal{Q}_B^T \mathcal{Q}_B = I$; $\mathcal{R}(\mathcal{Q}_A) = \mathcal{R}(A)$, $\mathcal{R}(\mathcal{Q}_B) = \mathcal{R}(B)$.
2. Compute the singular value decomposition (SVD): $\mathcal{Q}_A^T \mathcal{Q}_B = UCV^T$.
3. Extract the diagonal elements of $C$: $C_{ii} = \cos\theta_i$, to obtain the canonical correlations $\{\cos\theta_1, ..., \cos\theta_\alpha\}$.
```

The principal angles are the basic ingredient of a number of well-defined Grassmannian distances between subspaces (Ye and Lim, 2016). Here I use the chordal distance given by:

$$d(\mathcal{A}, \mathcal{B}) = \sqrt{\sum_{j=1}^{\alpha} \sin^2 \theta_j}. \tag{6.1}$$

The larger the chordal distance $d(\mathcal{A}, \mathcal{B})$ is, the worse the alignment between the subspaces $\mathcal{A}$ and $\mathcal{B}$.

I remark that the last inequality in $\alpha \leq \beta < n$ is strict. If a subspace spans the whole ambient space (i.e., $\beta = n$), then its distance to all other strict subspaces of $\mathbb{R}^n$ is trivially zero, as it is always possible to find a rotation that aligns the strict subspace with the whole space.

Alignment metric

My task involves establishing the alignment between three subspaces associated with the features X, the graph $\widehat{A}$, and the ground truth Y. To do so, I consider the distance matrix containing all the pairwise chordal distances:

$$D(X, \widehat{A}, Y) = \begin{bmatrix} 0 & d(X, \widehat{A}) & d(X, Y) \\ d(X, \widehat{A}) & 0 & d(\widehat{A}, Y) \\ d(X, Y) & d(\widehat{A}, Y) & 0 \end{bmatrix}, \tag{6.2}$$

and I take the Frobenius norm (Golub and Van Loan, 2013) of this matrix D as my subspace alignment measure (SAM):

$$\mathcal{S}(X, \widehat{A}, Y) = \|D(X, \widehat{A}, Y)\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{3} \sum_{j=1}^{3} D_{ij}^2}. \tag{6.3}$$

The larger $\|D\|_{\mathrm{F}}$ is, the worse the alignment between the three subspaces. This alignment measure has a geometric interpretation related to the area of the triangle with sides $d(X, \widehat{A}), d(X, Y), d(\widehat{A}, Y)$ (blue triangle in Figure 6.1).

Determining the dimension of the subspaces

The feature, graph, and ground truth matrices $(X, \widehat{A}, Y)$ are associated with subspaces of the ambient space $\mathbb{R}^N$, where N is the number of nodes (or samples). These subspaces are spanned by: the eigenvectors of $\widehat{A}$, the principal components of the feature matrix X, and the principal components of the ground truth matrix Y, respectively (Von Luxburg, 2007). The dimension of the graph subspace is N; the dimension of the feature subspace is the number of features $C^0 < N$ (in my examples); and the dimension of the ground truth subspace is the number of classes $F < C^0 < N$.

The pairwise chordal distances D_{ij} in Equation (6.2) are computed from a number of minimal angles, corresponding to the smaller of the two dimensions of the subspaces being compared. Hence the dimensions of the subspaces $(k_X, k_{\widehat{A}}, k_Y)$ need to be defined to compute the distance matrix D. Here, I am interested in finding low dimensional subspaces of features, graph and ground truth with dimensions $(k_X^*, k_{\widehat{A}}^*, k_Y^*)$ such that they provide maximum discriminatory power between the original problem and the fully randomised (null) model. To do this, I propose the following criterion:

$$\begin{aligned} k_Y^* &= F \\ (k_X^*, k_{\widehat{A}}^*) &= \arg\max_{k_X, k_{\widehat{A}}} \left(\|D(X_{100}, \widehat{A}_{100}, Y)\|_{\mathrm{F}} - \|D(X, \widehat{A}, Y)\|_{\mathrm{F}} \right). \end{aligned} \tag{6.4}$$

I choose k_Y^* equal to the number of ground truth classes since they are non-overlapping (Von Luxburg, 2007). My optimisation selects k_X^* and $k_{\widehat{A}}^*$ such that the difference in alignment between the original problem with no randomisation ($p_X = p_{\widehat{A}} = 0$) and an ensemble of 100 fully randomised (feature and graph, $p_X = p_{\widehat{A}} = 100$) problems is maximised (see SI for details on the optimisation scheme). This criterion maximises the range of values that $||D||_{\mathrm{F}}$ can take, thus augmenting the discriminatory power of the alignment measure when finding the alignment between both data sources and the ground truth, beyond what is expected purely at random. Importantly, the reduced dimension of features and graph are found simultaneously since my objective is to quantify the alignment (or amount of shared information) contained in the three subspaces. My criterion effectively amounts to finding the dimensions of the subspaces that maximise a difference in the surfaces of the blue and red triangles in Figure 6.1.

I provide the code to compute my proposed alignment measure at `https://github.com/haczqyf/gcn-data-alignment`.

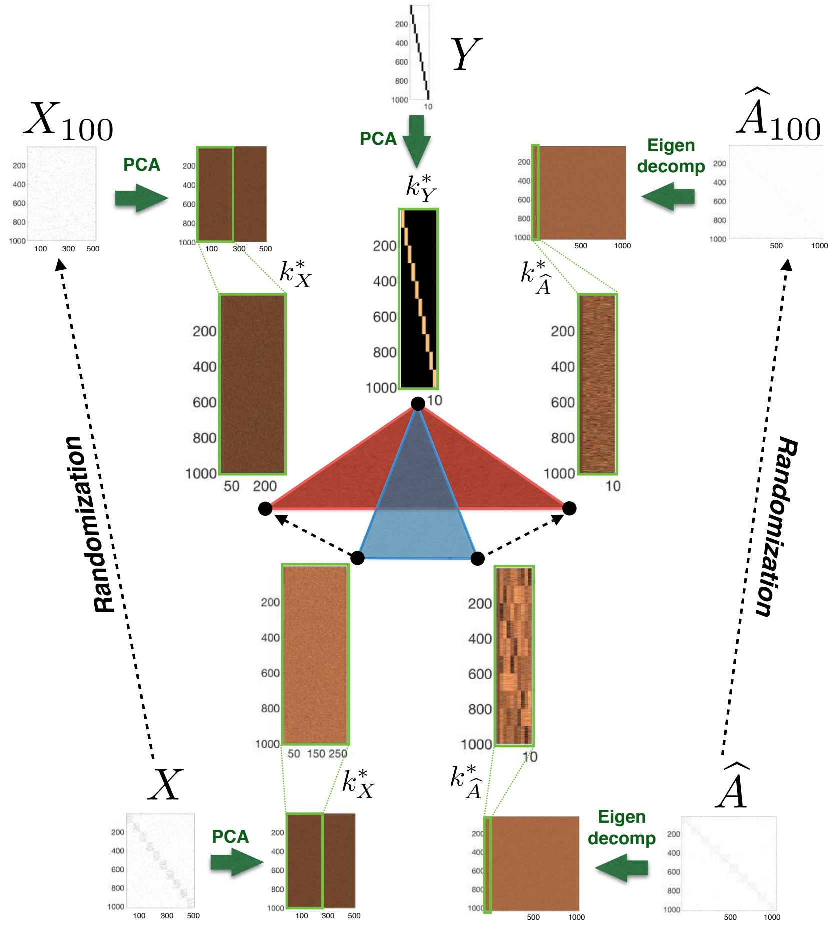


Figure 6.1: **Method to determine relevant subspaces (Equation (6.4)).** Using the constructive example, I illustrate the subspaces representing features, graph and ground truth. The feature and ground truth matrices are decomposed via PCA and the graph matrix is similarly eigendecomposed. Fixing $k_Y^* = F$, I optimise Equation (6.4) to find the dimensions k_X^* and $k_{\widehat{A}}^*$ that maximise the difference between the area of the blue triangle, which reflects the alignment of the three subspaces $(X, \widehat{A}, Y)$ of the original data, and the area of the red triangle, which corresponds to the alignment of the subspaces $(X_{100}, \widehat{A}_{100}, Y)$ of the fully randomised data. The edges of the triangles correspond to the pairwise chordal distances (e.g., the base of the blue triangle corresponds to $d(X, \widehat{A})$).

6.3 Experiments

6.3.1 Data sets

Relevant statistics of the data sets, including number of nodes and edges, dimension of feature vectors, and number of ground truth classes, are reported in Table 6.1.

Table 6.1: **Some statistics of the data sets in my study.**

Data sets	**Nodes** (N)	**Edges**	**Features** (C^0)	**Classes** (F)
Constructive	$1,000$	$6,541$	500	10
Cora	$2,485$	$5,069$	$1,433$	7
AMiner	$2,072$	$4,299$	500	7
Wikipedia	$20,525$	$215,056$	100	12
Wikipedia I	$2,414$	$8,163$	100	5
Wikipedia II	$1,858$	$8,444$	100	5

Constructive example

To illustrate the alignment measure in a controlled setting, I build a constructive example, consisting of $1,000$ nodes assigned to 10 planted communities $C_1, ..., C_{10}$ of equal size. I then generate both a feature matrix and a graph matrix whose structures are aligned with the ground truth assignment matrix. The graph structure is generated using a stochastic block model that reproduces the ground truth structure with some noise: two nodes are connected with a probability $p_{in} = 0.07$ if they belong to the same community C_i and $p_{out} = 0.007$ otherwise. The feature matrix is constructed in a similar way. The feature vectors are 500 dimensional and binary, i.e., a node either possesses a feature or does not. Each ground truth cluster is associated with 50 features that are present with a probability of $p_{in} = 0.07$. Each node also has a probability $p_{out} = 0.007$ of possessing each feature characterising other clusters. Using the same stochastic block structure for both features and graph ensures that they are maximally aligned with the ground truth. This constructive example is then randomised in a controlled way to detect the loss of alignment and the impact this loss of alignment has on the classification performance.

Cora

The Cora data set is a benchmark for classification algorithms using text and citation data[1]. Each paper is labelled as belonging to one of 7 categories (Case_Based, Genetic_Algorithms, Neural_Networks, Probabilistic_Methods, Reinforcement_Learning, Rule_Learning, and Theory), which gives the ground truth Y. The text of each paper is described by a 0/1 vector indicating the absence/presence of words in a dictionary of $1,433$ unique words, the dimension of the feature space. The feature matrix X is made from these word vectors. I extracted the largest connected component of this citation graph (undirected) to form the graph adjacency matrix A.

AMiner

For additional comparisons, I produced a new data set with similar characteristics to Cora from the academic citation site AMiner. AMiner is a popular scholarly social network service for research purposes only (Tang et al., 2008), which provides an open database[2] with more than 10 data sets encompassing researchers, conferences, and publication data. Among these, the academic so-

[1] https://linqs.soe.ucsc.edu/data

[2] https://aminer.org/data

cial network[3] is the largest one and includes information on papers, citations, authors, and scientific collaborations. In 2012 the Chinese Computer Federation (CCF) released a catalogue including 10 subfields of computer science. Using the AMiner academic social network, Ref. (Qian et al., 2017) extracted $102,887$ papers published from 2010 to 2012, and mapped each paper with a unique subfield of computer science according to the publication venue. Here, I use these assigned categories as the ground truth for a classification task. Using all the papers in Ref. (Qian et al., 2017) that have both abstract and references, I created a data set of similar size to Cora. I extracted the largest connected component from the citation network of all papers in 7 subfields (Computer systems/high performance computing, Computer networks, Network/information security, Software engineering/software/programming language, Databases/data mining/information retrieval, Theoretical computer science, and Computer graphics/multimedia) from 2010 to 2011. The resulting AMiner citation network consists of $2,072$ papers with $4,299$ edges. Just as with Cora, I treat the citations as undirected edges, and obtain an adjacency matrix A. I further extracted the most frequent 500 stemmed terms from the corpus of abstracts of papers and constructed the feature matrix X for AMiner using bag-of-words.

Wikipedia

As a contrasting example, I produced three data sets from the English Wikipedia. The Wikipedia provides an interlinked corpus of documents (articles) in different fields, which "cite" each other via hyperlinks. I first constructed a large corpus of

[3] https://aminer.org/aminernetwork

articles, consisting of a mixture of popular and random pages so as to obtain a balanced data set. I retrieved the $5,000$ most accessed articles during the week before the construction of the data set (July 2017), and an additional $20,000$ documents at random using the Wikipedia built-in random function[4]. The text and subcategories of each document, together with the names of documents connected to it, were obtained using the Python library Wikipedia[5]. A few documents (e.g., those with no subcategories) were filtered out during this process. I constructed the citation network of the documents retrieved and extracted the largest connected component. The resulting citation network contained $20,525$ nodes and $215,056$ edges. The text content of each document was converted into a bag-of-words representation based on the 100 most frequent words. To establish the ground truth, I used 12 categories from the API (People, Geography, Culture, Society, History, Nature, Sports, Technology, Health, Religion, Mathematics, Philosophy) and assigned each document to one of them. As part of my investigation, I split this large Wikipedia data set into two smaller subsets of non-overlapping categories: Wikipedia I, consisting of Health, Mathematics, Nature, Sports, and Technology; and Wikipedia II, with the categories Culture, Geography, History, Society, and People.

All six data sets used here can be found at `https://github.com/haczqyf/gcn-data-alignment/tree/master/alignment/data`.

[4] https://en.wikipedia.org/wiki/Wikipedia:Random

[5] https://github.com/goldsmith/Wikipedia

6.3.2 GCN architecture, hyperparameters and implementation

I used the GCN architecture (Kipf and Welling, 2017) and implementation[6] provided by the authors of (Kipf and Welling, 2017), and followed closely their experimental setup to train and test the GCN on my data sets. I used a two-layer GCN as described in Section 5.3 with the maximum number of training iterations (epochs) set to 400 (Kingma and Ba, 2015), a learning rate of 0.01, and early stopping with a window size of 100, i.e., training stops if the validation loss does not decrease for 100 consecutive epochs. Other hyperparameters used were: (i) dropout rate: 0.5; (ii) L2 regularisation: 5×10^{-4}; and (iii) number of hidden units: 16. I initialised the weights as described in Ref. (Glorot and Bengio, 2010), and accordingly L1 row-normalised the input feature vectors. For the training, validation and test of the GCN, I used the following split: (i) 5% of instances as training set; (ii) 10% as validation set; and (iii) the remaining 85% as test set. I used this split for all data sets with the exception of the full Wikipedia data set, where I used: (i) 3.5% of instances as training set; (ii) 11.5% as validation set; and (iii) the remaining 85% as the test set. This modification of the split was necessary to ensure the instances in the training set were evenly distributed across categories.

[6] https://github.com/tkipf/gcn

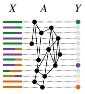
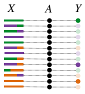
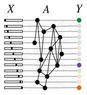
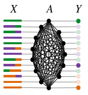
6.4 Results

The GCN performance is evaluated using the standard classification accuracy defined as the proportion of nodes correctly classified in the test set.

6.4.1 GCN: original graph vs. limiting cases

For each data set in Table 6.1, I trained and tested a GCN with the original graph and features matrices, and GCN models under the three limiting cases described in Section 6.2.2. I computed the average accuracy of 100 runs with random weight initialisations (Table 6.2).

Table 6.2: **Classification accuracy of GCN and limiting cases for my data sets.** The best performance is indicated in bold. Error bars are evaluated over 100 runs. The GCN with original data performs best in most cases, but is outperformed by MLP in the full Wikipedia data set and its subset Wikipedia II.

	GCN (original)	**GCN (limiting cases)**		
		No graph = MLP *(Only features)* $A = \mathbf{0}\mathbf{0}^T$	No features *(Only graph)* $X = I_N$	Complete graph *(Mean field)* $A = \mathbf{1}\mathbf{1}^T - I_N$
				
Data sets				
Constructive	**0.932** ± **0.006**	0.416 ± 0.010	0.764 ± 0.009	0.100 ± 0.003
Cora	**0.811** ± **0.005**	0.548 ± 0.014	0.691 ± 0.006	0.121 ± 0.066
AMiner	**0.748** ± **0.005**	0.547 ± 0.013	0.591 ± 0.006	0.123 ± 0.045
Wikipedia	0.392 ± 0.010	**0.450** ± **0.007**	0.254 ± 0.037	O.O.M.
Wikipedia I	**0.861** ± **0.006**	0.796 ± 0.005	0.824 ± 0.003	0.163 ± 0.135
Wikipedia II	0.566 ± 0.021	**0.659** ± **0.011**	0.347 ± 0.012	0.155 ± 0.176

The GCN using all the information available in the features and the graph outperforms MLP (the no graph limit) except in the case of the large Wikipedia set. Hence using the additional information contained in the graph does not necessarily increase the performance of GCN. To investigate this issue further, I split the Wikipedia data set into two subsets: Wikipedia I, with articles in topics that tend to be more self-referential (e.g., Mathematics or Technology) and Wikipedia II, containing pages in areas that are less self-contained (e.g., Culture or Society). If adjacent nodes tend to belong to the same class (i.e., being self-referential) in the citation graph, it can be understood that graph and ground truth are well aligned. I observed that GCN outperforms MLP for Wikipedia I but the opposite is still true for Wikipedia II. Finally, I also observe that the performance of "No features" is always lower than the performance of GCN, and, as expected, the performance of "Complete graph" (i.e., mean field) is very low and close to pure chance (i.e., $\sim 1/F$).

6.4.2 Performance of GCN under randomisation

The results above lead us to pose the hypothesis that a degree of synergy between features, graph and ground truth is needed for GCN to perform well. To investigate this hypothesis, I use the randomisation schemes described in Section 6.2.1 to degrade systematically the information content of the graph and/or the features in my data sets. Figure 6.2 presents the performance of the GCN as a function of the percent of randomisation of the graph structure, the features, or both. As expected, the accuracy decreases for all data sets as the information contained in the graph, features or both is scrambled, yet with differences in the decay rate of each of the ingredients for the different examples.

Note that the chance-level performance of the "Complete graph" (mean field) limiting case is achieved only when both graph and features are fully randomised, whereas the accuracy of the two other limiting cases ('No graph - MLP', "No features") is reached around the half-point ($\sim$ 50%) of randomisation of the graph or of the features, respectively. This indicates that using the scrambled information above a certain degree of randomisation becomes more detrimental to the classification performance than simply ignoring it.

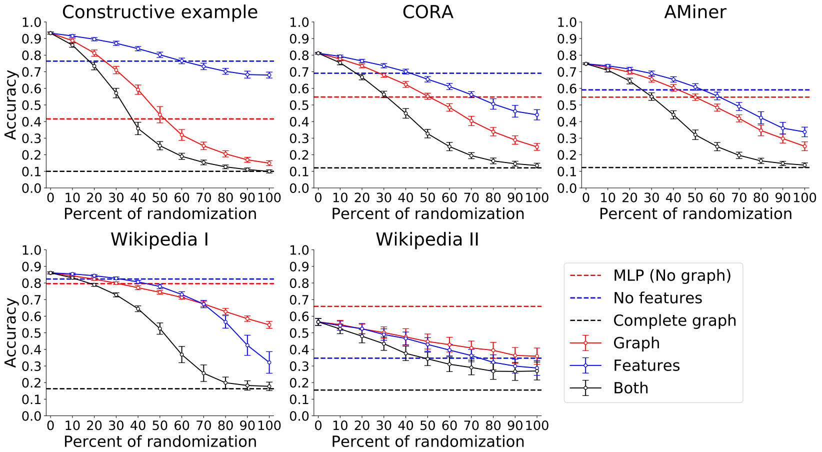


Figure 6.2: **Degradation of classification performance as a function of randomisation.** Each panel shows the degradation of the classification accuracy as a function of the randomisation of graph, features and both, for a different data set. Error bars are evaluated over 100 realisations: for zero percent randomisation, I report 100 runs with random weight initialisations; for the rest, I report 1 run with random weight initialisations for 100 random realisations. The horizontal lines correspond to the limiting cases in Table 6.2. The full Wikipedia data set was not analysed here since the eigendecomposition of $\widehat{A}$ needed to obtain $k^*_X, k^*_{\widehat{A}}$ is computationally intensive.

6.4.3 Relating GCN performance and subspace alignment

I tested whether the degradation of GCN performance is linked to the increased misalignment of features, graph and ground truth given by the SAM:

$$\mathcal{S}^*(X, \widehat{A}, Y) = \|D(X, \widehat{A}, Y; k^*_X, k^*_{\widehat{A}}, k^*_Y)\|_{\mathrm{F}} \tag{6.5}$$

which corresponds to Equation (6.3) computed with the dimensions $(k^*_X, k^*_{\widehat{A}}, k^*_Y)$ obtained using Equation (6.4) (Table 6.3, and see Section A.1 for the optimisation scheme used). Figure 6.3 shows that the GCN accuracy is clearly (anti)correlated with the subspace alignment distance (Equation (6.5)) in all my examples (mean correlation $= -0.92$). As I randomise the graph and/or features, the subspace misalignment increases and the GCN performance decreases. In addition to the Chordal distance, Ref. (Ye and Lim, 2016) studies other subspace distances. While all the distances can be expressed in terms of the principal angles θ_j, some rely on all the angles, whereas others only use the maximum principal angle. I obtain similar results for distances that use all the principal angles (e.g., Chordal, Grassmann), but I find that extremal distances based on the maximum principal angle (e.g., the Projection distance) do not correlate as well with GCN performance. This highlights the importance of the information captured by all principal angles to quantify the alignment between subspaces. For results based on the Grassmann and Projection distances, see Section A.3.

Table 6.3: **Dimensions of the three subspaces obtained according to Equation (6.4) for my data sets.**

Data sets	k^*_X	$k^*_{\hat{A}}$	k^*_Y
Constructive example	287	10	10
Cora	1,291	190	7
AMiner	500	57	7
Wikipedia I	68	1,699	5
Wikipedia II	100	1,125	5

6.5 Discussion and conclusion

In this chapter, I have introduced SAM (Equation (6.5)), a measure that quantifies the consistency between the feature and graph ingredients of data sets, and I showed that it correlates well with the classification performance of GCNs. My experiments show that a degree of alignment is needed for a GCN approach to be beneficial and that using a GCN can actually be detrimental to the classification

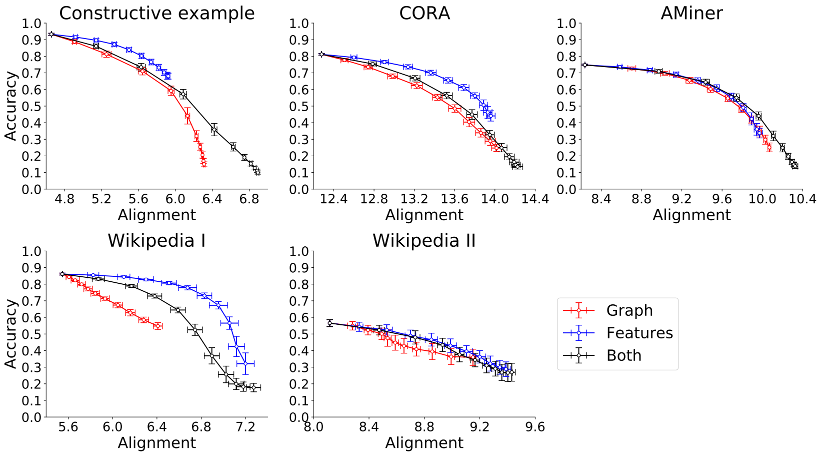


Figure 6.3: **Classification performance versus the SAM.** Each panel shows the accuracy of GCN versus the SAM (Equation (6.5)) for all the runs presented in Figure 6.2. Error bars are evaluated over 100 randomisations.

performance if the feature and graph subspaces associated with the data are not constructively aligned (e.g., Wikipedia and Wikipedia II).

6.5.1 Implications for research

My first set of experiments (Table 6.2) reflects the varying amount of information that GCN can extract from features, graph, and their combination, for the purpose of classification. For a classifier to perform well, it is necessary to find (possibly nonlinear) combinations of features that map differentially and distinctively onto the categories of the ground truth. The larger the difference (or distance on the projected space) between the samples of each category, the easier it is to "separate" them, and the better the classifier. In the MLP setting, for instance, the weights between layers (W^ℓ) are trained to maximise this separation. As seen by the different accuracies in the "No graph" column (Table 6.2), the features of each example contain a variable amount of information that is mappable on its ground truth. A similar reasoning applies to classification based on graph information alone, but in this case, it is the eigenvectors of $\widehat{A}$ that need to be combined to produce distinguishing features between the categories in the ground truth (e.g., if the graph substructures across scales (Lambiotte et al., 2014) do not map onto the separation lines of the ground truth categories, then the classification performance based on the graph will deteriorate). The accuracy in the "No features" column indicates that some of the graphs contain more congruent information with the ground truth than others. Therefore, the "No graph" and "No features" limiting cases inform about the relative congruence of each type of information with respect to the ground truth. One can then conjecture that if the performance of the "No features" case is higher than the "No graph" case, GCN will yield better results than MLP. These results suggest that in future work researchers should consider limiting cases to understand the potential contribution of features and graph to the classification performance of GCN.

In addition, my numerics show that although combining both sources of information generally leads to improved classification performance ('GCN original' column in Table 6.2), this is not always necessarily the case. Indeed, for the Wikipedia and Wikipedia II examples, the classification performance of the MLP ('No graph'), which is agnostic to relationships between samples, is better than when the additional layer of relational information about the samples (i.e., the graph) is incorporated via the GCN architecture. This suggests that, for improved GCN classification, the information contained in features and graph needs to be constructively aligned with the ground truth. This phenomenon can be intuitively understood as follows. In the absence of a graph (i.e., the MLP setting), the training of the layer weights is done independently over the samples, without assuming any relationship between them. In GCN, on the other hand, the role of the graph is to guide the training of the weights by averaging the features of a node with those of its graph neighbours. The underlying assumption is that the relationships represented by the graph should be consistent with the information of their features, i.e., the features of nodes that are graph neighbours are expected to be more similar than otherwise; hence the training process is biased towards convolving the diffusing information on the graph to extract improved feature descriptions for the classifier. However, if feature similarities and graph neighbourhoods (or more generally, graph communities (Lambiotte et al., 2014)) are not congruent, this graph-based averaging during the training is not beneficial. These findings provide theoretical contributions to the design of proper GNN frameworks that take the similarity of node features into account in the graph convolution layers.

To explore this issue in a controlled fashion, my second set of experiments (Figure 6.2) studied the degradation of the classification performance induced by the systematic randomisation of graph structure and/or features. The erosion of information is not uniform across my examples, reflecting the relative salience of each of the components (features and graph) for classification. Note that the GCN is able to leverage the information present in any of the two components and is only degraded to chance-level performance when both graph and features are fully randomised. Interestingly, this fully randomised (chance-level) performance coincides with that of the “Complete graph” (or mean field) limiting case, where the classifier is trained on features averaged over all the samples, thus leading to a uniform representation that has zero discriminating power when it comes to category assignment.

These results suggest that a degree of constructive alignment between the matrices of features, graph, and ground truth $(X, \widehat{A}, Y)$ is necessary for GCN to operate successfully beyond standard classifiers. To capture this idea, I proposed a simple SAM (Equation (6.5)) that uses the minimal principal angles to capture the consistency of pairwise projections between subspaces. Figure 6.3 shows that SAM correlates well with the classification performance and captures the monotonic dependence remarkably, given that SAM is a simple linear measure being applied to the outcome of a highly non-linear, optimised system. The results are consistent for other versions of GCN. In particular, in Section A.2 I show that the alignment measure correlates well with the performance of the recently proposed Simple Graph Convolution (SGC) (Wu et al., 2019). This new simple measure offers new perspectives to the currently highly debated problem (McCabe et al., 2021) of how to measure the consistency or distance between different sources of information (e.g., features or graph).

6.5.2 Implications for practice

The proposed alignment measure can be used to measure the consistency between the graph, features, and ground truth and thus indicates the potential classification performance of GCN. For example, I applied this measure to show why certain graphs are good for classification when multiple graphs are given as candidates in Chapter 7 (see Section 7.4.2).

The alignment measure can also be used to evaluate the relative importance of features and graph for classification without explicitly running the GCN, by comparing the SAM under full randomisation of features against the SAM under full randomisation of the graph. If $\mathcal{S}^*(X_{100}, \widehat{A}, Y) > \mathcal{S}^*(X, \widehat{A}_{100}, Y)$, the features play a more important role in GCN classification. Conversely, if $\mathcal{S}^*(X_{100}, \widehat{A}, Y) < \mathcal{S}^*(X, \widehat{A}_{100}, Y)$, the graph is more important in GCN classification.

More generally, the SAM has potentially a wide range of applications in the quantification of data alignment, including, among others: quantifying the alignment of different graphs associated with, or obtained from, particular data sets; evaluating the quality of classifications found using unsupervised methods; and aiding in choosing the classifier architecture that is computationally most advantageous for a particular data set.

6.5.3 Limitations

My approach has a number of limitations that could be addressed in future work. First, it contains two parameters (i.e., the dimensions of the subspaces, k_X^* and $k_{\widehat{A}}^*$) which need to be tuned through a computational search. Second, the alignment is not directly comparable across data sets since the subspace dimensions are adjusted for each data set. To facilitate comparisons across data sets, normalised versions of the alignment measure will be the object of future work. Third, the current measure is not suitable for very large data sets as the eigendecomposition of large matrices is computationally demanding. For very large data sets, approximations (e.g., using the Lanczos algorithm to explore only leading eigenvectors) might be necessary to optimise the subspace dimensions. While I have focused here on node classification, it would be interesting in future work to extend my measure to other tasks such as graph classification, link prediction, and regression.

6.6 Contribution to the literature

My work is concerned with the conceptual understanding of GCN, a prominent deep learning architecture in the research area of machine learning with graphs. My study has supported the hypothesis that a certain degree of alignment among the graph, feature, and ground truth is needed in order to make GCN perform well. To measure the consistency between the data ingredients in GCN, I introduced a SAM with spectral and geometric interpretations to quantify the alignment among data ingredients and showcased that it correlates well with the performance of GCN.

My findings are particularly timely given the increasing interest of the scientific community in incorporating relational data into classification tasks. More generally, SAM can be used to inform the choice and development of the most advantageous architecture that takes into account the degree of alignment between ingredients and is applicable to other situations where the alignment between data sets, or between graphs, or between graphs and data needs to be computed.

Chapter 7

Geometric graphs from data to aid classification

7.1 Introduction

Recently, work with GCNs (Kipf and Welling, 2017) has suggested that using a graph of samples in conjunction with sample features can improve classification performance when compared with traditional methods that use only features. Computationally, the graph allows the definition of a convolution operation that exchanges and aggregates the features of samples that are connected on the graph. If the graph and the features align well with the underlying class structure as suggested in Chapter 6 and in Ref. (Qian et al., 2021b), the graph convolution operation homogenises features of neighbouring nodes. These neighbouring nodes will tend to be more similar.

In many instances, extra relational information in the form of a graph is not readily available. However, the intuition that nodes that are close in feature space tend to belong to the same class can still be exploited by constructing geometric graphs directly from the data features, and in doing so, creating neighbourhoods of similar samples. Such feature-derived graphs can then be used to aid and potentially sharpen the classification.

Here, I explore the benefit of constructing geometric graphs from the features of the samples and using them within a GCN for sample classification (Figure 7.1a). Graph construction, or inference, is a problem encountered in many fields (Newman, 2018a), from neuroimaging to genetics, and can be based on many different types of heuristics, from simple thresholding (Zalesky et al., 2012) or statistically significant group-level thresholding (Lord et al., 2012) to sophisticated regularisation schemes (Omranian et al., 2016). In general, the goal is to obtain graphs that concisely preserve key properties of the original data set as sparsely as possible, i.e., with a low density of edges.

In this work, I use several popular geometric graph constructions to extract graphs from data and study how the classification performance depends on the graph construction method and the edge density. I find that there is a range of relatively low edge densities over which the constructed graphs improve the classification performance. Among the construction methods, I show that the recently proposed Continuous k-Nearest Neighbour (CkNN) (Berry and Sauer, 2019) performs best for GCN classification.

To gain further intuition about the role played by the graph in improving classification, I compute two simple measures: (i) the alignment of the convolution of graph and features with the ground truth; and (ii) the ratio of class separation in the output activations of the GCN. I show that the optimised geometric graphs increase the alignment and the class separation. Finally, I show that the graphs can be made more efficient using spectral graph sparsification (Spielman and Srivastava, 2011), which reduces the edge density of the optimised CkNN graphs while improving further the classification performance.

7.2 Methods

7.2.1 Graph construction

Let X_i be the F-dimensional feature vector (L1-normalised) of the i-th sample of a data set with N samples. The pairwise dissimilarity between samples i and j is taken to be the Euclidean distance:

$$d(i,j) = \|X_i - X_j\|_2 \, . \tag{7.1}$$

The distance matrix of all samples $D \in R^{N \times N}$ where $D_{ij} = d(i,j)$ is then used to construct unweighted and undirected graphs based on different heuristics. To guarantee connectedness over the data set, I first construct the Minimum Spanning Tree (MST). The MST is obtained from the Euclidean distance matrix D using the Kruskal algorithm and contains the $N-1$ edges that connect all the nodes (samples) in the graph with a minimal sum of edge weights (distances). Once the weighted MST is obtained, I ignore the edge weights, as is also done for all other graphs in this chapter. Thus the resulting graphs are undirected and unweighted. The 0-1 adjacency matrix of the MST is denoted by A^{MST}. I then add edges to the MST based on two types of criteria: (i) local neighbourhoods or (ii) balancing local and global distances.

Methods Based on Local Neighbourhoods: Nearest Neighbours

The objective of neighbourhood-based methods is to construct a sparse graph by connecting two samples if they are local neighbours, as determined by $d(i, j)$.

The simplest such algorithm is k-Nearest Neighbour (kNN). A kNN graph has an edge between two samples i and j if one of them belongs to the k-nearest neighbours of the other. The adjacency matrix $A^{\text{kNN}} \in R^{N \times N}$ of a kNN graph is defined by:

$$A_{i,j}^{\text{kNN}} = \begin{cases} 1 & \text{if } d(i,j) \leq d(i,i_k) \text{ or } d(i,j) \leq d(j,j_k) \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

where i_k and j_k represent the k-th nearest neighbours of samples i and j, respectively.

Although widely used, kNN has limitations. Perhaps most importantly, kNN graphs can have highly heterogeneous degree distributions and often contain hubs, i.e., samples with a high number of connections, since kNN greedily connects two samples as long as one of them belongs to the other's k-nearest neighbours. It has been suggested that the presence of hubs in kNN graphs is particularly severe when the samples are high-dimensional (Radovanović et al., 2010). It has been observed that hubs tend to deteriorate the classification accuracy of semi-supervised learning (Ozaki et al., 2011).

To overcome this limitation, the Mutual k-Nearest Neighbour (MkNN) algorithm, a variant of kNN, was proposed (Ozaki et al., 2011). In an MkNN graph an edge is established between samples i and j if each of them belongs to the other's k-nearest neighbours. The adjacency matrix $A^{\text{MkNN}} \in R^{N \times N}$ of the MkNN graph is defined by:

$$A_{i,j}^{\text{MkNN}} = \begin{cases} 1 & \text{if } d(i,j) \leq d(i,i_k) \text{ and } d(i,j) \leq d(j,j_k) \\ 0 & \text{otherwise} \end{cases} \tag{7.3}$$

Note that the MkNN algorithm guarantees that the degrees of all samples are bounded by k. Therefore, MkNN reduces the presence of hubs when k is adequately small.

Another limitation of kNN is its lack of flexibility to provide a useful, stable graph when the data is not uniformly sampled over the underlying space, which is often the case in practice (Liu and Barahona, 2020). In such situations, it is difficult to find a single value of k that can accommodate the disparate levels of sampling density across the data since the kNN graph will connect samples with very disparate levels of similarity depending on the region of the sample space (i.e., in densely sampled regions, the graph only connects data points that are very similar, whereas, in poorly sampled regions, the graph connects data samples that can be quite dissimilar). This disparity biases the training of the GCN. The non-uniformity of the data distribution thus makes it difficult to tune a unique k parameter that is appropriate across the whole data set. If the value of k is too small, the graph is dominated by local noise and fails to provide consistent information to improve the GCN training. If the value of k is large, the resulting graph is over-connected and leads GCN to degraded accuracy, close to mean-field classification. Hence, when the sampling is not homogeneous, standard kNN graphs can be sub-optimal in capturing the underlying data structure with a view to improved learning.

CkNN (Berry and Sauer, 2019) has recently been introduced to address this limitation by allowing an adjusted local density. The adjacency matrix $A^{\text{CkNN}} \in R^{N\times N}$ associated with a CkNN graph is defined by:

$$A_{i,j}^{\text{CkNN}} = \begin{cases} 1 & \text{if } d(i,j) < \delta\sqrt{d(i,i_k)d(j,j_k)} \\ 0 & \text{otherwise} \end{cases} \tag{7.4}$$

where the parameter $\delta > 0$ regulates the density of the graph. For a fixed k, the larger δ is, the denser the CkNN graph becomes. Ref. (Berry and Sauer, 2019) shows that the CkNN graph captures the geometric features of the data set with the additional consistency that the unnormalised Laplacian of the CkNN graph converges spectrally to the Laplace-Beltrami operator in the limit of large data. In this work, I fix $\delta = 1$ and vary k so that CkNN can be compared with kNN and MkNN, as suggested in Ref. (Liu and Barahona, 2020).

All these three methods capture the geometry of local neighbourhoods, with global connectivity guaranteed by the MST.

Balancing Local and Global Distances: Relaxed Minimum Spanning Tree

Alternatively, other graph constructions attempt to balance the local geometry with a measure of global geometry extracted from the full data set. In recent years, several algorithms have been introduced to explore global properties of the data using the MST (Beguerisse-Díaz et al., 2013; Liu and Barahona, 2020). Here, I focus on the Relaxed Minimum Spanning Tree (RMST) (Beguerisse-Díaz et al., 2013), which considers the largest distance $d^{\max}_{\text{MST-path}(i,j)}$ encountered along the unique MST-path between i and j. If $d^{\max}_{\text{MST-path}(i,j)}$ is substantially smaller than $d(i,j)$, RMST discards the direct link between i and j, recognising the multi-step MST-path as a good model to capture the similarity between them. If, on the other hand, $d(i,j)$ is comparable to $d^{\max}_{\text{MST-path}(i,j)}$, the MST-path does not provide a good model, and RMST adds the direct link between i and j. The adjacency matrix $A^{\text{RMST}} \in R^{N\times N}$ associated with a RMST graph is defined by:

$$A^{\text{RMST}}_{i,j} = \begin{cases} 1 & \text{if } d(i,j) < d^{\max}_{\text{MST-path}(i,j)} + \gamma(d(i,i_k) + d(j,j_k)) \\ 0 & \text{otherwise} \end{cases} \tag{7.5}$$

where $\gamma \geq 0$ is the density parameter, and $d(i,i_k)$ and $d(j,j_k)$ approximate the local distribution of samples around i and j, respectively, as the distance to their k-th nearest neighbour (Zemel and Carreira-Perpiñán, 2005). Here, I fix $k = 1$ and vary γ to change the edge density, as in Ref. (Liu and Barahona, 2020).

7.2.2 Spectral graph sparsification

The graph construction methods above can be thought of as graph densification, in which the starting point is the MST over the N samples and an edge is added between two samples i and j if the distance $d(i, j)$ meets a defined criterion. Graph sparsification operates in the opposite direction: starting from a given graph, the goal is to obtain a sparser graph that approximates the original graph so that it can be used, e.g., in numerical computations, without introducing too much error. Sparsified graphs are more efficient for both numerical computation and data storage (Spielman and Teng, 2011). Here, I focus on spectral graph sparsification (Spielman and Teng, 2011), and apply the seminal Spielman-Srivastava sparsification algorithm (SSSA) proposed in Ref. (Spielman and Srivastava, 2011). SSSA obtains a spectral approximation of the given graph that satisfies the following criterion:

$$(1-\sigma)\, x^{\mathrm{T}} L x \leq x^{\mathrm{T}} \widetilde{L} x \leq (1+\sigma)\, x^{\mathrm{T}} L x \tag{7.6}$$

where $x \in R^{N\times 1}$ is a node vector, and L and $\widetilde{L}$ are the Laplacian matrices of the original and sparsified graphs, respectively.

7.3 Experiments

7.3.1 Data sets

I use seven data sets collected from various sources. I provide the data sets at `https://github.com/haczqyf/ggc/tree/master/ggc/data`. The data set statistics are summarised in Table 7.1.

1. I consider three data sets *Constructive*, *Cora* and *AMiner* as introduced in Section 6.3.1.

2. *Digits* is a handwritten digits data set. Each sample is an 8x8 image of a digit. This is one of the benchmark data sets for classification in Scikit-learn (Pedregosa et al., 2011).

3. *FMA*: The original data set (Defferrard et al., 2017; Franceschi et al., 2019) contains 140 audio features extracted from $7,994$ music tracks. I use this data set to address the problem of genre classification. The original data set in ref. (Franceschi et al., 2019) contains 8 genres. I randomly sample $2,000$ music tracks (250 for each genre) to produce my data set.

4. *Cell*: This is a data set of brain cell types from autism. The original data set (Velmeshev et al., 2019) contains the gene expression values ($\log_2$ transformed 10x UMI counts from cellranger) of $104,599$ single cells from brains of control individuals and of patients with autism, where each cell (sample) is characterised by the expression level of $36,501$ genes (features). The full data set contains cells from 17 cell types (categories). To produce

my data set, I randomly sample $2,000$ cells from 10 cell types (200 cells for each type) and select as my features the expression level of the top 500 most highly variable genes across the $2,000$ cells in my sample.

5. *Segmentation*: This is an image segmentation data set, which is provided at UCI machine learning repository (Dua and Graff, 2017) at `https://archive.ics.uci.edu/ml/datasets/Image+Segmentation`. Each sample represents an image described by 19 high-level and man-crafted numeric-valued attributes.

Table 7.1: Summary statistics of the data sets in my study.

Data sets	**Type**	**Samples (N)**	**Features (F)**	**Classes (C)**	**Train/Validation/Test**
Constructive	Stochastic block model	$1,000$	500	10	50/100/850
Cora	Text (Bag-of-words)	$2,485$	$1,433$	7	$119/253/2,113$
AMiner	Text (Bag-of-words)	$2,072$	500	7	$98/212/1,762$
Digits	Images (Grayscale pixels)	$1,797$	64	10	$80/189/1,528$
FMA (songs)	Music track features	$2,000$	140	8	$96/204/1,700$
Brain cell types	Single-cell transcriptomics	$2,000$	500	10	$100/200/1,700$
Segmentation	Image features	$2,310$	19	7	$112/234/1,964$

7.3.2 Graph construction

I consider geometric graph constructions that fall broadly in two groups: (i) three methods based on local neighbourhoods, i.e., kNN, MkNN and CkNN (Berry and Sauer, 2019) graphs; and (ii) a method that balances local and global distances measured on the MST, i.e., the RMST (Beguerisse-Díaz et al., 2013). In all cases, I start from an MST to guarantee the resulting graph comprises a single connected component, and I add edges based on the corresponding distance heuristics. In each construction, a parameter regulates the edge density of the graph: k in kNN, MkNN, and CkNN, and γ in RMST.

For each data set and each graph construction, I find the edge density that maximises the average GCN classification accuracy on the validation set by sweeping over 50 values of the edge density, from sparse to dense. Note that this is a hyperparameter optimisation process, rather than exact optimisation in the mathematical sense. For each value of the density, I run the GCN classifier 10 times starting from random weight initialisations to compute the average accuracy. Note that the two limiting cases are well characterised: the "no graph" limit corresponds to the MLP; the "complete graph" limit is equivalent to the mean field and leads to random class assignment (Qian et al., 2021b).

7.3.3 Graph sparsification

Sparse graphs are generally favoured over dense graphs, particularly for large data sets, as they are more efficient for numerical computation and data storage. I investigate whether it is possible to sparsify the optimised geometric graphs obtained above while preserving, or even improving, GCN classification performance. Motivated by the key importance of spectral properties in graph partitioning (Delvenne et al., 2010; Lambiotte et al., 2014), I apply the SSSA (Spielman and Srivastava, 2011) to my optimised CkNN graphs. The SSSA reduces the number of edges of a graph while preserving the spectral content of the graph Laplacian given by Equation (7.6).

I apply the SSSA to the optimised CkNN and select the sparsification that maximises the classification accuracy on the validation set. For each data set, I obtain increasingly sparse versions of the optimised geometric graph computed above by scanning over 50 values of the sparsity parameter σ between $1/N$ and 1. At each value of σ, I run the GCN classifier 10 times starting from random weight initialisations and compute the average accuracy over the validation set. I then select the graph with the highest accuracy and maximum sparsity. If sparsification does not improve performance on the test set, I report the unsparsified graph as optimal (e.g., in Cora, Digits, and FMA in Figure 7.3b).

7.3.4 GCN architecture, hyperparameters and implementation

My study applies the two-layer GCN described in Section 5.3. I use the GCN implementation provided by the authors of Ref. (Kipf and Welling, 2017), and follow closely the experimental setup in Refs. (Kipf and Welling, 2017; Qian et al., 2021b). I use a two-layer GCN with $2,000$ epochs (training iterations); learning rate of 0.01; and early stopping with a window size of 200. Other hyperparameters are: dropout rate: 0.5; L2 regularisation: 5×10^{-4}; number of hidden units: 16. The weights are initialised as described in Ref. (Glorot and Bengio, 2010), and the input feature vectors are L1 row-normalised. I choose the same data set splits as in Ref. (Qian et al., 2021b) with 5% of samples as training set, 10% of samples as validation set, and the remaining 85% as test set (see Table 7.1). The samples in the training set are evenly distributed across classes.

7.3.5 Graph-less classification methods

For comparison, I consider four graph-less classification methods: (i) MLP, which is equivalent to GCN with no graph (Kipf and Welling, 2017; Qian et al., 2021b); (ii) kNN Classification (kNNC) based on the plurality vote of its k-nearest neighbours; (iii) Support Vector Machine (SVM) with Radial Basis Function kernel; and (iv) Random Forest (RF). I use the Scikit-learn (Pedregosa et al., 2011) implementation for kNNC, SVM, and RF. For each method and each data set, I use the validation set to optimise the following hyperparameters: number of neighbours (kNNC); regularisation parameter (SVM); maximum depth (RF). All other hyperparameters are set as default in Scikit-learn. I compare the graph-less methods against the MLP = GCN (No graph) used as the reference baseline.

7.3.6 Data and code availability

I provide the data sets and code for geometric graph construction at `https://github.com/haczqyf/ggc`. The code for GCNs is provided by the authors of (Kipf and Welling, 2017) at `https://github.com/tkipf/gcn`. The code for kNNC, SVM, and RF can be found at `https://scikit-learn.org/stable/` from scikit-learn (Pedregosa et al., 2011). The code for SSSA is available at `https://epfl-lts2.github.io/gspbox-html/doc/utils/gsp_graph_sparsify.html` from Graph Signal Processing Toolbox (Perraudin et al., 2014).

7.3.7 Algorithm complexity

For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes $v_i \in \mathcal{V}$ and $|\mathcal{E}|$ edges $(v_i, v_j) \in \mathcal{E}$, the time complexity for GCN, i.e., to evaluate Equation (9), is $O(|\mathcal{E}|FHC)$ (Kipf and Welling, 2017), where $|\mathcal{E}|$ is the number of graph edges, F is the dimension of the feature space, H is the number of units in the hidden layer and C is the number of classes in the ground truth. Hence the computational complexity for GCN is linear in the number of graph edges. For the geometric graph construction, a brute force approach to computing exactly a geometric graph (i.e., the kNN-type graphs) has time complexity $O(FN^2)$. However, fast approximate kNN graph algorithms were proposed to reduce this time complexity. I mention two examples: (i) Ref. (Chen et al., 2009) proposes an algorithm with complexity $O(FN^t)$ with $1 < t < 2$, and (ii) Ref. (Andoni and Indyk, 2006) proposes an algorithm that uses locality sensitive hashing, which has complexity $O\left(FN^{1/c^2+o(1)}\right)$ where $c = 1 + \epsilon > 1$. For a thorough list of approximate kNN algorithms, see https://github.com/stephenleo/adventures-with-ann/. Regarding spectral sparsification, the algorithm is nearly linear with time complexity $\tilde{O}(|\mathcal{E}|)$ (Spielman and Srivastava, 2011), where the $\tilde{O}$ notation ignores logarithmic factors. Finally, for the MST construction, I use the Kruskal algorithm implemented in Scipy with time complexity $O(|\mathcal{E}|\log N)$.

7.3.8 Run time and memory requirements

To give a sense of run times and memory requirements for my algorithm, I summarise the numbers briefly for the Cora data set, which presents the worstcase run times and storage requirements among my seven examples. Indeed, I find that Cora has the longest run times, consistent with the algorithmic complexity introduced in Section 7.3.7, since Cora has the largest number of nodes and highest dimensions. For graph construction, creating and storing in disk all the graphs during the optimisation of the hyperparameter takes around 13 hours with a maximum used memory of around 3G. However, my algorithm can be further optimised since the graphs do not have to be stored and could be created and used on the fly to save memory usage and access time. Furthermore, over-dense graphs could be avoided altogether since the optimised graphs usually are relatively sparse. Indeed, I find that the graphs with optimal accuracy have densities on the order of $0.005 - 0.05$ of the total number of possible edges (see Table B.2), and for densities above ~ 0.1 the accuracy drops below the accuracy of an MLP. For higher densities, the accuracy consistently degrades towards the random assignment limit. Therefore the grid search of the hyperparameter can be restricted to low-density graphs, and dense graphs do not have to be stored or computed. The search for the optimal hyperparameter can be further aided with a bisection scheme and could be parallelised to improve the efficiency of the optimisation.

For each value of the hyperparameter, I run a GCN 10 times from 10 random initialisations. The cost of each GCN is moderate: the complexity of GCN scales nearly linearly with the number of edges of the graph. The cost of constructing kNN-type graphs (originally of $O(N^2)$) can also be reduced to nearly linear in the number of nodes with approximation algorithms. Sparsification is also nearly linear, as shown by Spielman. Hence the methodology has the potential to be applied to relatively large graphs with further code optimisation. For instance, each GCN for Cora typically takes less than 7 minutes for relatively sparse graphs ($k \leq 200$), and each graph sparsification takes less than 2 minutes.

Comparing to the Louvain-based methods, there is the same complexity for the kNN graph construction, whereas the run time complexity of Louvain optimisation is $O(N\log^2 N)$. For Seurat, there is the additional cost of performing PCA to extract the top T principal components, with complexity $O(N^2T)$ (inherited from randomised SVD). Thus, the run time complexity and memory requirements of the Louvain-based methods are comparable to those of my method.

7.4 Results

7.4.1 Geometric graphs constructed from data features can aid sample classification

Figure 7.1b shows the classification performance of a GCN with a CkNN graph of increasing density applied to a data set of computer science papers (AMiner), which I use as my running example throughout. I find that adding relatively sparse graphs improves the classification accuracy, reaching a maximum increase of 10.9% at an edge density of 0.039 ($k^* = 199$) on the validation set. Once the edge density parameter is optimised on the validation set, I apply the GCN classifier to the test set, and the test accuracy is recorded. In this case, the GCN yields an improvement of 7.2% in classification accuracy on the test set with respect to MLP, as reported in Table 7.2.

I have investigated six real-world data sets from different domains, ranging from text (AMiner (Qian et al., 2017; Tang et al., 2008), Cora (Sen et al., 2008)) to music track features (FMA (Defferrard et al., 2017; Franceschi et al., 2019)) to single-cell transcriptomics (Cell (Velmeshev et al., 2019)) to imaging (Digits (Pedregosa et al., 2011), Segmentation (Dua and Graff, 2017)). I have also studied one constructive data set with a well-defined ground truth based on a stochastic block model. For a detailed description of the data sets, see Section 7.3.1. I have compared the performance of four graph-less, feature-based classifiers (MLP, kNNC, SVM, and RF) to GCN classifiers with optimised feature-derived geometric graphs (Table 7.2). My numerical experiments indicate that the GCNs with feature-derived graphs generally achieve better classification performance than graph-less classifiers. In particular, the CkNN graph construction achieves the highest accuracy improvement (+8.3% on average above MLP) across my seven data sets.

Table 7.2: Classification accuracy (in percent) on the test set (averaged over 10 runs with random initialisations) for 7 data sets with 8 classifiers (four graph-less methods; GCN with four graph constructions). The standard deviation is reported in Table B.1. The top two results for each data set are bold. Overall, GCN with CkNN graphs displays the best performance. The density parameters of optimised graphs are reported in Table B.2.

Classifier	Constructive	Cora	AMiner	Digits	FMA	Cell	Segmentation	Average improvement
MLP = GCN (No graph)	42.1	54.2	54.4	82.0	34.3	79.5	72.0	–
kNNC	31.4	38.2	28.0	88.3	30.6	58.7	68.8	(−10.6)
SVM	40.0	55.9	51.4	87.7	35.3	81.5	**87.7**	(+3.0)
RF	36.3	56.1	47.7	83.0	33.0	**88.0**	**88.8**	(+2.1)
GCN (kNN)	**53.9**	**66.4**	59.2	92.0	**35.6**	83.8	83.5	(+8.0)
GCN (MkNN)	45.2	64.1	**61.8**	**93.2**	**35.6**	84.0	83.0	(+6.9)
GCN (CkNN)	**51.1**	**66.6**	**61.6**	**93.4**	**36.0**	84.0	83.9	(+8.3)
GCN (RMST)	45.9	64.8	61.5	89.3	35.4	**84.9**	83.0	(+6.6)

7.4.2 The role of feature-derived graphs in classification

My results show improved classification performance of GCNs with feature-derived geometric graphs of appropriate edge density. Indeed, over-sparse graphs perform close to MLPs, the "no graph" limiting case, whereas over-dense graphs are

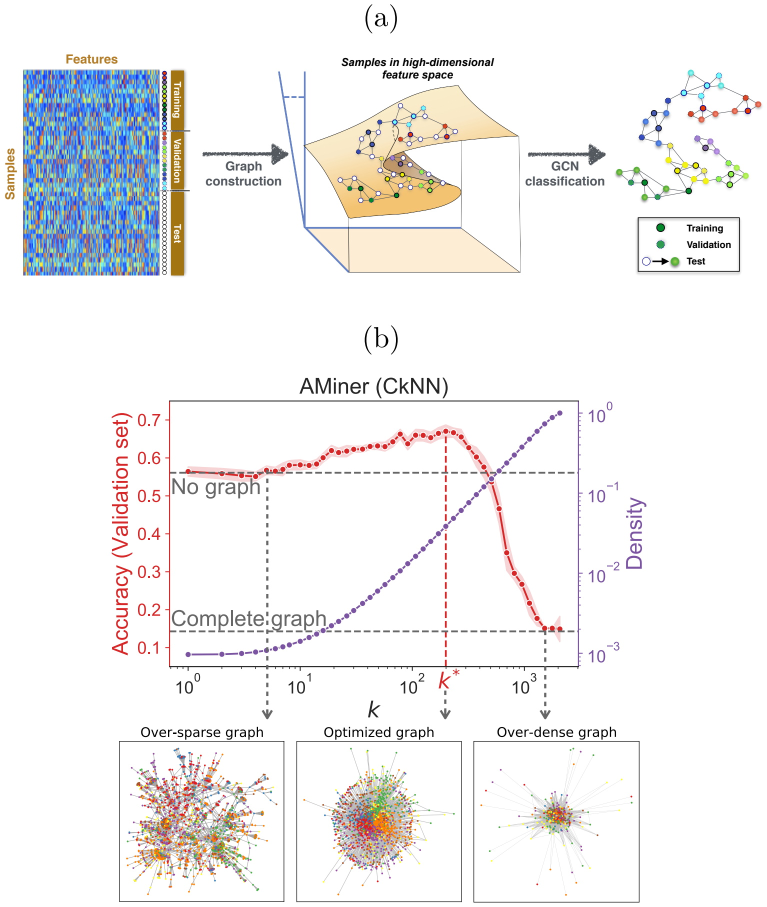


Figure 7.1: **Geometric graphs constructed from data features can aid sample classification.**

(a) Workflow for GCN classification using feature-derived graphs. (b) The validation set is used to search for graphs with optimised edge density—here illustrated with the AMiner data set and CkNN graph construction. In red, the GCN classification accuracy on the validation set as a function of the density parameter, k. The results are averaged over 10 runs with random weight initialisations; shaded region represents standard deviation. As I sweep k from "no graph" (MLP) to complete graph (mean field, random assignment), the classification accuracy on the validation set exhibits a maximum for a CkNN graph with density parameter k^*. In purple, edge density of the CkNN graphs as k is varied. Figures for all graph constructions and data sets are provided in Figure B.1. Also shown below, graph visualisations using the spring layout for over-sparse, optimised and over-dense graphs, with nodes coloured according to their ground truth class.

detrimental, as they approach the “mean field” limit that behaves like a random class assignment. Hence there is a sweet spot of relatively low edge density where graphs improve the performance maximally. To gather further insight into the role of the constructed graphs in classification, I explore their properties from two complementary perspectives.

Over-dense graphs degrade the alignment of graph-convolved features with the ground truth

Consider the classification of N samples with F features into C classes making use of a graph with adjacency matrix A. In Chapter 6 and Ref. (Qian et al., 2021b) it was shown that good GCN performance requires a certain degree of alignment between the linear subspaces associated with the matrix of features, $X \in R^{N\times F}$, the adjacency matrix of the graph with self-loops, $\widehat{A} \in R^{N\times N}$, and the ground truth membership matrix, $Y \in R^{N\times C}$. Inspired by Chapter 6 and Ref. (Qian et al., 2021b), I evaluate the alignment between the ground truth Y and the graph-convolved features $X_A := \widehat{A}X$ as:

$$S(X, \widehat{A}, Y) = \cos(\theta_1(\mathcal{X}_{\mathcal{A}}, \mathcal{Y})). \tag{7.7}$$

Here $\theta_1(\mathcal{X}_{\mathcal{A}}, \mathcal{Y})$ is the minimal principal angle (Björck and Golub, 1973; Golub and Van Loan, 2013; Knyazev and Argentati, 2002) between the column spaces of the matrices $\mathrm{PCA}(X_A, p^*)$ and $\mathrm{PCA}(Y, p^*)$, which contain the top principal components, as determined by the parameter p^*, of $\widehat{A}X$ and Y, respectively. The parameter p^* is the ratio of explained variance that maximises the Pearson

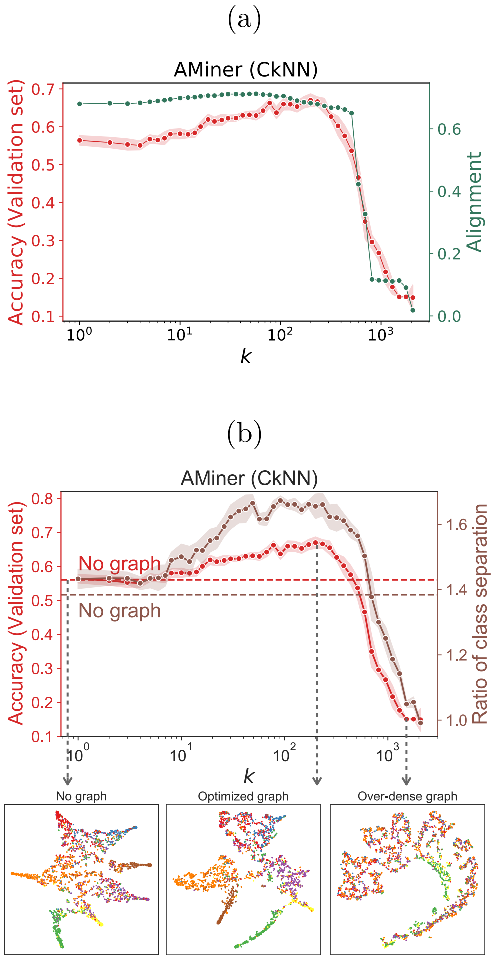


Figure 7.2: The role of feature-derived graphs in classification. (a) In green, I show the alignment (Equation (7.7)) of CkNN graphs for the AMiner data set as a function of the density parameter k. In red, classification accuracy as in Figure 7.1b. The drop in classification accuracy corresponds to the drop in the subspace alignment. Results for all graph constructions and data sets are given in Figure B.2. (b) Ratio of class separation (Equation (7.8)) computed from the output activations of the GCN with CkNN graphs for AMiner data set as a function of the density parameter k, in brown. The results are averaged over 10 runs with random weight initialisations; shaded region is the standard deviation. The brown dashed line represents the RCS for the MLP, i.e., GCN with no graph. In red, classification accuracy, as in Figure 7.1b. Below, I show two-dimensional t-SNE projections of the output activations of GCNs with no graph, optimised graph and over-dense graph. The nodes are coloured according to the ground truth class labels. The optimised graph induces higher class separability, as shown by an increased RCS and better resolved t-SNE projection. Results for all graph constructions and data sets are provided in Figure B.3.

correlation between the alignment (Equation (7.7)) and the classification accuracy on the validation set.

Figure 7.2a shows the alignment (Equation (7.7)) between the ground truth and the graph-convolved data for CkNN graphs of increasing density on the AMiner data set. I find that the reduction in classification accuracy induced by over-dense graphs is linked to a strong disruption of the subspace alignment $S\left(X, \widehat{A}, Y\right)$. In the limit of the complete graph, the alignment approaches the value of 0, i.e., the minimal angle $\theta_1 = \pi/2$, indicating that the two subspaces are orthogonal. On the other hand, Sparse graphs induce a slight increase of the subspace alignment at the same time as improving the classification accuracy. The alignment and classification accuracy show a good correlation for the AMiner data set: the Pearson correlation between alignment and accuracy (validation set) is 0.970, obtained for a value of $p^* = 0.4$. The same procedure has been carried out for all seven data sets, and the results are presented in Figure B.2. The Pearson correlation coefficient between alignment and accuracy (validation set) ranges from 0.602 (Segmentation) to 0.970 (AMiner) with an average of 0.852 over all 7 data sets, thus indicating a good correspondence between the classification accuracy and the graph-induced alignment of data and ground truth.

Graphs with optimised density increase the ratio of class separation

Another way of assessing the effect of the constructed graphs on classification is to study the inherent separability of the probabilistic GCN assignment matrix, i.e., the row-stochastic matrix $Z \in R^{N \times C}$ of output activations in Equation (5.10). The effect of the graph on Z reflects the quality of the classifier: a good graph should enhance the separation of samples from different classes while clustering together samples from the same class in C-dimensional space. I quantify the separability of the GCN mapping using $Z' \in R^{N\times 2}$, the two-dimensional t-SNE (Maaten and Hinton, 2008) embedding of Z, on which I compute the ratio between the average inter-class and intra-class distances, denoted ratio of class separation (RCS):

$$\text{RCS} = \frac{(\mathbf{1}^T (D^{(Z')} \circ M^{\text{inter}})\mathbf{1})/(\mathbf{1}^T M^{\text{inter}}\mathbf{1})}{\left(\mathbf{1}^T \left(D^{(Z')} \circ M^{\text{intra}}\right) \mathbf{1}\right)/(\mathbf{1}^T M^{\text{intra}}\mathbf{1})}. \tag{7.8}$$

Here, $D^{(Z')}$ is the Euclidean distance matrix for the t-SNE embedding Z', i.e., $D_{ij}^{(Z')} = \left\| Z_i' - Z_j' \right\|_2$; the notation $\circ$ represents the Hadamard, element-wise, matrix product; $M^{\text{inter}} \in R^{N\times N}$ is the inter-class indicator matrix, i.e., $M_{ij}^{\text{inter}} = 1$ if samples i and j belong to different classes and $M_{ij}^{\text{inter}} = 0$ otherwise; and, conversely, $M^{\text{intra}} \in R^{N\times N}$ is the intra-class indicator matrix. Compactly, I have

$$M^{\text{inter}} = \mathbf{1}\mathbf{1}^T - YY^T$$
$$M^{\text{intra}} = YY^T - I_N,$$

where $I_N \in R^{N\times N}$ is the identity matrix and $\mathbf{1}$ is the N-dimensional vector of ones.

Figure 7.2b shows the RCS (Equation (7.8)) computed from the output activation of GCNs with CkNN graphs of increasing density (AMiner data set). I observe a high correlation between RCS and the classification accuracy (validation set): the Pearson correlation coefficient for AMiner is 0.953. Similar figures for all data sets are shown in Figure B.3. The Pearson correlation coefficient between RCS and accuracy (validation set) is high for all data sets, ranging from 0.876

(Segmentation) to 0.976 (Cora), with an average Pearson correlation coefficient of 0.938 across all 7 data sets. These results indicate that sparse graphs unfold the data and facilitate class separation, as illustrated by the t-SNE plots and the increased RCS; on the other hand, over-dense graphs reduce separability and eventually converge to the mean field limiting value of RCS $= 1$, i.e., when there is no distinction between inter- and intra-class separation.

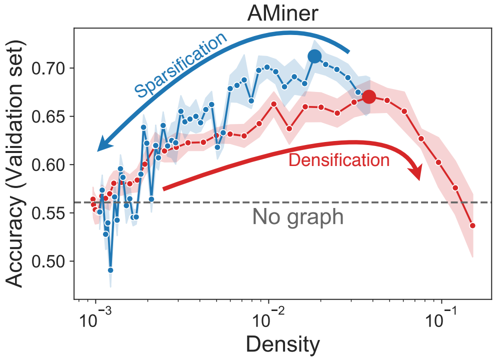
7.4.3 Spectral sparsification of optimised geometric graphs can further improve classification

Figure 7.3a shows that for the AMiner data set, it is possible to improve the classification accuracy using sparser graphs obtained with SSSA. This procedure was repeated for all seven data sets (see Figure B.4). For several of my data sets, the sparsified graphs perform better on the test data with reduced edge density (see Table 7.3b). The results of the sparsification are robust: starting the sparsification from three different highly-optimised CkNN graphs leads to similar results (see Figure B.4 and Table B.3). Furthermore, the sparsification induces increased alignment and RCS, which correlates with the improved classification accuracy on the validation set (see Figures B.5 and B.6).

7.5 Discussion and conclusion

Supervised classification assigns unseen samples to classes based on their features by learning from examples with known class labels. I show that classification can

(a)



(b)

	Optimised CkNN		**Sparsification of optimised CkNN**	
Data set	⟨Degree⟩	Accuracy (Test)	⟨Degree⟩	Accuracy (Test)
Constructive	9.2	51.1	6.3	51.6
Cora	36.7	66.6	36.7	66.6
AMiner	79.8	61.6	38.1	62.5
Digits	28.1	93.4	28.1	93.4
FMA	8.0	36.0	8.0	36.0
Cell	15.0	84.0	4.8	85.0
Segmentation	10.3	83.9	8.2	84.0
Average improvement		(+8.3)		(+8.7)

Figure 7.3: Spectral sparsification of optimised geometric graphs can further improve classification. (a) In red, the same data as in Figure 7.1b, i.e., classification accuracy of GCN with CkNN graphs on AMiner data set for increasing edge density; 10 runs with random weight initialisations, shaded area is standard deviation. The large red dot indicates the optimised graph found as edges are added (densification). Starting from this optimised graph, I reduce the number of edges using the SSSA (sparsification) and record the classification accuracy on the validation set, in blue; 10 runs with random weight initialisations, shaded region is standard deviation. The large blue dot indicates the optimised sparsified graph. The grey dashed line corresponds to the classification accuracy of the MLP (no graph) on the validation set. Results for all data sets are provided in Figure B.4. (b) Comparison of optimised graphs obtained through the densification and sparsification processes. The average degree of the graph (⟨Degree⟩) and classification accuracy in percent on the test set are reported; averaged over 10 runs with random weight initialisations. Overall, sparsified graphs exhibit improved accuracy on the test set with lower edge density.

be improved by using the sample features not only as the basis for classification, but also as a means to construct geometric graphs that encapsulate the closeness between samples. Such feature-derived graphs can be used within graph-based deep-learning models to improve the classification. To understand the benefits of these graphs, I show that they align the data to the class labels and enhance class separability. I also demonstrate how to make the graphs sparser and hence more efficient while still potentially improving their performance.

7.5.1 Implications for research

My empirical study used data sets from different domains to show that sparse geometric graphs constructed from data features can aid classification tasks when used within the framework of GNNs. It is worth noting that although here I have used the widely popular GCN framework to perform the classification task, other advanced GNN architectures (e.g., Deep Graph Infomax (Veličković et al., 2019)) could be incorporated in my pipeline as an alternative to GCN for this purpose. In my numerics, GCN with CkNN geometric graphs display the largest improvement in classification accuracy (Table 7.2). This result is in line with recent work on geometric graph construction for data clustering (Liu and Barahona, 2020), which showed improved behaviour of CkNN over other neighbourhood methods, such as kNN. CkNN graphs have been recently proposed as a consistent discrete approximation of the Laplace-Beltrami operator governing the diffusion on an underlying manifold (Berry and Sauer, 2019). Since GCN uses the graph to guide the diffusion of features to neighbouring nodes, this offers a natural explanation for the good performance of CkNN under the GCN framework. RMST graphs use a criterion that balances neighbourhood distances with non-local distances in the data set within my graph construction methods. While RMST outperforms graph-less methods, it does not outperform neighbourhood-based methods in the examples considered here. However, RMST graphs could be appropriate for data sets where similarities based on longer paths are important. For other researchers in future work, these observations will provide guidance on how to select the most appropriate graph construction methods for aiding classification.

Intuitively, geometric graphs capture the closeness (i.e., similarity) between samples in feature space and can thus be helpful to learn and channel class labels from known samples to similar unseen samples. To gain further insight into why geometric graphs can improve GCN classification performance, I showed that the graph induces an increased alignment of features and ground truth, as measured by the simple measure (Equation (7.7)). The alignment correlates well with classification performance, specifically capturing the deleterious effect that over-dense graphs have on classification performance (Figure 7.2a). When the graphs are over-dense, they lead to a "mean field" averaging over the whole data set, which breaks the alignment—an analogous problem to the over-smoothing observed when there are too many layers in GCNs (Chen et al., 2020; Li et al., 2018a). I also showed that graphs with appropriate density induce increased class separability, as measured by the ratio of class separation (RCS, Equation (7.8)) derived from the GCN output activations, whereas over-sparse and over-dense graphs lead to lower class separability (Figure 7.2b). These two simple measures (i.e., alignment and RCS) contribute to the current debate on how to explain the functionality of GNNs (Ying et al., 2019). Deviating from strictly geometric graphs, I demonstrated that spectral sparsification (SSSA) applied to the optimised CkNN graphs can be used to reduce the number of edges whilst still improving the classification performance (Figure 7.3). My choice of a spectral criterion for sparsification stems from the fact that the preserved Laplacian quadratic form (Equation (7.6)) is strongly related to graph partitioning and community detection (Delvenne et al., 2010; Lambiotte et al., 2014). The resulting efficient graphs are thus the product of a mixed process: a geometric graph provides a local similarity neighbourhood which is further sharpened using global graph properties captured by the Laplacian spectrum. Researchers can benefit from these findings by incorporating graph sparsification algorithms to design more efficient GNN architectures.

Methods that leverage graphs in data analysis have a long history (Lauritzen, 1996), and have been recently considered in conjunction with deep learning algorithms. Ref. (Franceschi et al., 2019) proposed a novel method that jointly learns graph structure and the parameters of a GNN by solving a bilevel program to obtain a discrete probability distribution on the edges of the graph. I have compared my method with the one proposed in Ref. (Franceschi et al., 2019). My results are summarised in Table B.4 and indicate that my proposed method achieves, on average, classification accuracy comparable to Ref. (Franceschi et al., 2019), yet with a significantly smaller number of parameters, thus simplifying the training and reducing the inclination to overfitting. Table B.4 also includes the results (Qian et al., 2021b) obtained by applying GCN to data sets that contain a graph as an additional source of information (i.e., the citation networks for Cora and AMiner). The improved accuracy of GCN with these original graphs stems from the additional information the graphs contain beyond what is present in the features alone. Specifically, the original graphs for Cora and AMiner collate citations between scientific articles, which encode additional information about the similarity between articles not captured by the features (i.e., the text embedding vectors) of the articles themselves. Another recent method constructed a local neighbourhood graph as part of convolution-based classifiers (Wang et al., 2019b). My work also contributes to this ongoing debate on how to leverage GCN when graphs are not available. In contrast to the above related works, I focused on graph-theoretical measures (Liu and Barahona, 2020), by exploring different graph constructions and the importance of edge density and spectral content for classification and characterising the effect of graphs through geometric notions of separability and subspace alignment (Qian et al., 2021b).

7.5.2 Implications for practice

In my numerical experiments, feature-derived geometric graphs appear to be most useful when the data is high-dimensional, noisy, and co-linearity is present in the features. In particular, GCN with optimised graphs outperforms the graph-less methods in all my data sets except "Segmentation". All my data sets are high-dimensional without feature engineering except "Segmentation", a data set with 19 engineered features specifically optimised for classification—this is the set-up where SVM and RF are expected to work well. However, even in that case, I note that the featured-derived graphs still improve the classification performance with respect to MLP, indicating that the graphs help filter out feature similarities that can obscure the action of MLP.

Beyond the potential to improve performance, using graphs to aid classification changes the paradigm from supervised to semi-supervised learning. Supervised methods, e.g., MLP, perform inductive learning, whereas graph-based semi-supervised learning can be either transductive or inductive. GCNs belong to transductive learning, since the graph of the full data set is used for the training. Therefore, whilst potentially advantageous, the use of GCNs can also restrict the generalisability to new samples. In many applications, such a requirement does not impose severe restrictions, but graph-based methods can still be adapted to classify new data without the need to recompute the model. For instance, one could predict the class label of a new sample directly from the output activations Z of the closest samples in the original set, or using more elaborate diffusion-based schemes (Peach et al., 2020).

My proposed pipeline also shares common ground with some of the most successful clustering methods developed for single-cell genomics data sets. For example, Seurat (Satija et al., 2015) uses Louvain modularity (Blondel et al., 2008) optimisation to perform community detection on a kNN graph constructed from the top principal components of data. Similarly, other methods for graph-based clustering have been introduced using multiscale extensions of the Louvain algorithm in the framework of Markov Stability (Liu and Barahona, 2020). Although classification and clustering are different learning tasks, I have carried out a comparison between my proposed method (CkNN+GCN) and two Louvain-based clustering methods (Seurat and a straightforward kNN+Louvain clustering). After optimising each method using the training and validation sets, I computed the assignment it produces on the test set and compared it to the ground truth classes (see Section B.1). The quality of the assignments (evaluated with the Adjusted

Rand Index and Normalised Mutual Information) presented in Table B.5 indicates that, on average, across my data sets, my proposed method performs better than Seurat's approach, and this could facilitate applications in related fields such as computational biology.

7.5.3 Limitations

There are some dimensions of choices of algorithms in this chapter that could be further explored. Here I explored graph construction based on geometry; it will be interesting to consider graph construction paradigms that incorporate other criteria, e.g., small-worlds (Watts and Strogatz, 1998), graph expanders (Hoory et al., 2006), or entangled networks (Donetti et al., 2005), among others. Similarly, although I showed that spectral sparsification (Spielman and Teng, 2011) is a good choice to improve efficiency, other graph sparsification paradigms, e.g., cut sparsification (Fung et al., 2019), might also be helpful to achieve efficient graphs for classification. Here I have adopted the Euclidean distance as a simple metric to base my geometric graph construction. However, other metrics could be used in my pipeline and could be indeed more appropriate for different types of data. Investigating the effect of different distance metrics (such as the Manhattan distance, cosine similarity, or distances in transformed spaces such as PCA, diffusion map, or other projections) would be an important question for future research. While I have focused here on the same hyperparameters of GCN as (Kipf and Welling, 2017), in future work it will be interesting to investigate whether different geometric graphs require different sets of hyperparameters to achieve the best performance.

7.6 Contribution to the literature

In this chapter, I studied the following research question: can we leverage GCN when no graphs exist explicitly in empirical domains? To this end, I systematically explored how feature-derived graphs, constructed and optimised in a simple and problem-agnostic manner, can be used with GCNs to improve classification performance on data sets where graphs are not originally available. This allows us to leverage recent graph-based deep-learning algorithms and extend the applicability of GNNs to feature-only data sets.

More specifically, I highlighted that: (i) Geometric graphs from data can be used in deep learning to improve classification; (ii) Optimised graphs align the data to the class labels and enhance class separability; (iii) Sparsifying the optimised graph can potentially improve classification performance; and (iv) Extensive experiments are performed on data sets from various scientific domains. My findings are particularly timely given the increasing interest within the machine learning community, and especially among scholars working on GNN, in combining graphs with classification and learning tasks.

Chapter 8

Conclusion

8.1 Brief overview

In the last two decades, networks have played an increasingly important role in multiple scientific domains, ranging from the social sciences to physics to computer science. In this thesis, I mainly focus on three types of networks–citation networks, social networks, and collaboration networks–by combining theories and methods from network science, sociology, machine learning, and data science. Specifically, I presented four projects concerned with two research clusters: social capital and deep learning. At a first glance, the two research clusters may seem to be only loosely connected, but in fact they can be unified by the underpinning idea that incorporating both graph structure and non-topological node features can provide more powerful representations of nodes than in cases where only one of them is leveraged.

The first project was presented in Chapter 3. Social capital extracted from network structures has long been seen as playing an important role in maintaining or hindering a wide range of performance-related outcomes at the individual and group levels. However, recent studies of social capital in networked communities have suggested that the network topology in itself is not sufficient in explaining how individuals, groups, and organisations can benefit from social interaction. Non-topological features at the node level need to be accounted for to better understand the performance implications of structure. To address this limitation, I developed new measures of network effective size, i.e., intra- and inter-brokerage, based on a certain non-topological property of nodes in directed and weighted networks, which can provide more fine-grained perspectives on social capital. I further obtained the corresponding simplified versions for undirected and unweighted networks, and derived the relationship between these two measures and the intra- and inter-local clustering coefficients. A case study on a co-authorship network showed that the new measures – intra- and inter-brokerage – can indeed capture distinct brokerage opportunities that would otherwise remain hidden by using standard brokerage measures.

The second project was presented in Chapter 4. This study aimed to explore the social capital of cities extracted from the collaboration patterns of their resident scientists and their external collaborators. To this end, by combining four large-scale bibliometric data sets, I started by constructing the scientific collaboration networks of city-identified scientists using 17 time windows over the period 1990 – 2006. I then quantified source of social capital (brokerage, strong ties, and diversity) and scientific performance (impact and innovation) of cities based on the collaboration network patterns of resident scientists and their external collaborators. This resulted in a panel data set containing 641 cities grouped into 64 countries. In the empirical regression analysis, I used three-level hierarchical random-intercept models. Results suggested that the relationship between the (internal or external) brokerage and scientific performance of cities is moderated by internal or external strong ties and the cities' geographical diversity.

The third project was presented in Chapter 6. I showed that the classification performance of GCNs is related to the alignment between features, graph, and ground truth, which I quantified using a subspace alignment measure corresponding to the Frobenius norm of the matrix of pairwise chordal distances between three subspaces associated with features, graph, and ground truth. The proposed measure is based on the principal angles between subspaces and has both spectral and geometrical interpretations. I showcased the relationship between the SAM and the classification performance through the study of limiting cases of GCNs and systematic randomisations of both features and graph structure applied to a constructive example and several examples of citation networks of different origins. The analysis also revealed the relative importance of the graph and features for classification purposes.

The fourth project was presented in Chapter 7. Traditional classification tasks learn to assign samples to given classes based solely on sample features. This paradigm is evolving to include other sources of information, such as known relations between samples. Here I showed that, even if additional relational information is not available in the data set, one can improve classification by constructing geometric graphs from the features themselves and using them within a Graph Convolutional Network. The improvement in classification accuracy is maximised by graphs that capture sample similarity with relatively low edge density. I showed that such feature-derived graphs increase the alignment of the data to the ground truth while improving class separation. I also demonstrated that the graphs can be made more efficient using spectral sparsification, which reduces the number of edges while still improving classification performance. I illustrated my findings using synthetic and real-world data sets from various scientific domains.

8.2 Contribution to the literature

This thesis leveraged various theories and methods from multiple research domains, including sociology, the science of science, machine learning, and network science. On the one hand, I developed an interdisciplinary approach to the study of citation, social, and collaboration networks. On the other hand, my work has made contributions to knowledge in each of these research domains, which I shall summarise in this section.

8.2.1 Sociology

My key contribution to sociology is to show that social capital can be quantified by combining both network structure and non-topological node features. This dovetails with current debates in the social sciences highlighting that most network-based theories of social capital tend to ignore the non-structural determinants of social capital. I proposed novel brokerage measures by extending Burt's effective size, and showed that they can provide finer-grained perspectives on social capital.

This will open new avenues for future research in the social sciences concerned with social capital considering there is an increasing interest in incorporating people's individual characteristics (e.g., gender and ethnicity) into social network analysis to reduce the potential issues of bias. In addition, my work on cities shows how the geo-social approach can contribute to measuring the social capital of cities, thus extending the scope of analysis from traditional observational units (such as people and organisations) to geographical places. Finally, my work concerned with deep learning demonstrates the powerful ability of learning representations with GNN methods. This can stimulate novel applications based on deep learning to sociological domains concerned with "learning" social capital in contrast to classic rule-based measures of social capital (e.g., effective size) developed by sociologists.

8.2.2 Science of science

Citation, collaboration, and social networks are essential types of networks that are studied and used in the science of science. In this case, my thesis, focusing on these networks, provides advancement of knowledge in the science of science from various aspects. First, the proposed intra- and inter-brokerage measures can be used to study how finer-grained brokerage based on a certain node attribute is associated with scientific success and thus informs scientists and policymakers on how to build social and collaboration networks in science. Second, I showed how to use interconnected geo-social network representations to provide new insights on the social capital of cities extracted from both resident scientists and external collaborators. In so doing, my work has policy implications for policy-makers on improving the scientific performance of a city considering both types of scientists.

Finally, citation networks are key examples I used in the two projects on deep learning. I demonstrated that both citation graph and node features (e.g., text data) can be used together to learn the embedding of scientific papers with advanced graph-based deep learning methods (e.g., GCNs). The learned node embeddings can then be used for downstream tasks, such as classification.

8.2.3 Intersection of machine learning and network science

My work concerned with deep learning lies at the intersection between machine learning and network science. As a junior network scientist, I engaged with the growing literature about machine learning with graph-structured data. My key contributions to the intersection between machine learning and network science are two-fold: (i) I showed that a certain degree of alignment between graph, features, and ground truth is needed for node classification using GCNs. This can allow machine learning practitioners to think about the role graph structure played in the learning process, given that there is a whole field called network science dedicated to studying the structures and dynamics of networks; and (ii) I demonstrated that, when there are no graphs available in a data set, one could use graph-theoretical approaches to infer and construct graphs that can aid machine learning tasks, e.g., classification. Scientists in these two communities of machine learning and network science could learn from, and contribute to, each other's field since there are overlapping research interests and shared methodologies.

8.3 Future work

I have been lucky to be offered a postdoctoral position at the Centre for Science of Science and Innovation[1] at the Kellogg School of Management at Northwestern University in the US.I envisage working on projects concerned with the science of science (Fortunato et al., 2018) and addressing several research questions that I shall briefly outline below. I will conduct my future studies by leveraging the expertise of network science and machine learning that I developed during my PhD.

8.3.1 Graph representation learning and science of science

Graphs are ubiquitous mathematical abstractions that can describe complex systems of relations and interactions. Science can be seen as a heterogeneous information network composed of scientists, publications, publication venues, topics, and interactions among them. The recent development on graph representation learning (Bronstein et al., 2017) in the machine learning community provides powerful computational methods, such as GNNs, that allow us to learn embeddings of node or graph level to investigate questions in the science of science by leveraging large-scale bibliometric data sets such as Microsoft Academic Graph. I propose to explore the following research avenues:

[1] https://www.kellogg.northwestern.edu/research/science-of-science.aspx

Journal representation learning

The goal is to combine the citation network among journals and journal-level high-dimensional features (e.g., topics that appeared in the journal) to learn low-dimensional representations of journals in an unsupervised manner with GNNs. The learned journal vector representations can be used in a number of applications: (i) quantifying the multidisciplinarity of journals through the pairwise distance between journals using the vector representations of journals; (ii) visualising science mapping; and many others.

Topic representation learning

With a similar learning framework as before but a focus on scientific topics instead of journals, the learned topic vector can be used to quantify the distance of topics, which could contribute to measuring the novelty of papers. For example, a paper associated with topics far away from each other can be considered as one with high novelty. This will further allow us to study the relationship between the novelty of papers and their future success or failure (Wang et al., 2019a).

8.3.2 Scientific career

The asymmetry of transition of scientists between research fields

Recent evidence shows that scientists' research interests keep evolving during their careers (Jia et al., 2017; Zeng et al., 2019). Although science tends to be more interdisciplinary, the transition between fields for scientists seems not symmetric.

For example, it appears that the transition from physics to social science is more likely than that from social science to physics. Can we find evidence for this hypothesis? Is there any topic hierarchy that facilitates or hinders the transition of a scientist? By addressing these questions, I will shed light on the possibility of career transition for young scientists. This will likely inform university leaders on the setting and arrangement of courses for students in less-advantaged fields towards interdisciplinary research. For example, students in the social sciences would benefit from learning computational methods with a view to building a career as computational social scientists and establishing stronger collaborations with scholars within the computational fields (e.g., computer science and physics).

8.3.3 Science of science for artificial intelligence research

The science of science uses large-scale data to search not only for universal patterns but also for domain-specific regularities. Artificial intelligence becomes increasingly important because it provides promisingly new tools to improve the intelligence of our society. Although a few SciSci studies have focused on AI (Frank et al., 2019), there is still relatively little comprehensive SciSci research dedicated to AI that considers its domain-specific characteristics. AI is different from traditional fields (e.g., physics and chemistry) in many ways. For example, unlike in other traditional fields, scholars in AI tend to publish papers in conferences as well as in journals, and pay large attention to the reproducibility of research results. Under the global technological innovation competition (e.g., between the US and China), it is essential to study the patterns and dynamics of collaboration, career, and citation for AI research at different levels, including scholars, institutions, and countries. The objective of my future study will be to understand the evolution of collaboration patterns, and then assess whether these network patterns are associated with scholars', institutions', and countries' performance.

8.3.4 Incorporating network representations beyond pairwise interactions into science of science

Recently, the network science community has showed an increasing interest in using network representations beyond pairwise interactions, including hypergraphs and simplicial complexes (Battiston et al., 2020; Lambiotte et al., 2019). These two representations offer a more realistic view to model real-world phenomena such as scientific collaboration. In the past few years, most studies at the intersection between the science of science and network science have used ordinary networks considering pairwise interactions, e.g., co-authorship networks and keyword co-appearance networks. These representations will inevitably filter out the higher-order relationships. A simple example is the case of a closed triangle in a co-authorship network, which cannot tell us whether the three connected scientists have indeed co-authored a common paper.

A handful of works have already made an initial effort and showed interesting results in the direction of incorporating higher-order network representations into the science of science (see Refs. (Patania et al., 2017; Salnikov et al., 2018)). Inspired by these previous works, I will leverage rich longitudinal bibliometric data sets and consider using tools and techniques developed by network scientists on hypergraphs or simplicial complexes to study various questions related to knowledge creation and diffusion.

Appendix A

Appendix of Chapter 6

A.1 Finding optimal dimensions

A key element of the SAM described in the main paper is to find lower dimensional representations of the graph, features and ground truth.

To determine the dimension of the representative subspaces, I propose the following heuristic:

$$(k_X^*, k_{\widehat{A}}^*) = \underset{k_X, k_{\widehat{A}}}{\arg\max} \left(\|D(X_{100}, \widehat{A}_{100}, Y)\|_{\mathrm{F}} - \|D(X, \widehat{A}, Y)\|_{\mathrm{F}} \right). \tag{A.1}$$

I choose k_Y^* to be equal to the number of categories in the ground truth as they are non-overlapping. Thus, k_X^* and $k_{\widehat{A}}^*$ range from k_Y^* to their maximum values, C^0, the dimension of the feature vectors, and N, the number of nodes in the graph, respectively.

To find the values for k_X^* and $k_{\widehat{A}}^*$, I scan different possible combinations of k_X and $k_{\widehat{A}}$. I applied two rounds of scanning. In the first scanning round, in the intervals of k_X and $k_{\widehat{A}}$, I picked 10 equally spaced values that contain the minimum and maximum possible values for k_X and $k_{\widehat{A}}$. For example, in Cora, k_Y^* equals 7 because the number of categories in the ground truth is 7. Thus k_X ranges from 7 to $1,433$. At the end of the first round, the optimal values of k_X^* and $k_{\widehat{A}}^*$ are $1,433$ and 282, respectively (see Figure A.1c).

In the second scanning round, I applied a very similar process to the one just described. I set the scanning intervals of k_X and $k_{\widehat{A}}$ as the neighbours of k_X^* and $k_{\widehat{A}}^*$ found in the first round, respectively. For example, in Cora, for the second round, I set the intervals of k_X and $k_{\widehat{A}}$ as $[1,274,1,433]$ and $[7,557]$. Again, I split the new intervals with 10 equally spaced values. I have also shown the scanning results for other data sets in Figure A.1.

A.2 Replicating experiments on a variant of GCN

First, I would like to highlight that the alignment metric is independent of the architecture and only relies on the data. Therefore, I expect that the conclusion will be consistent with different variants of GCNs: the convolution operation in GCN (Kipf and Welling, 2017) can be seen as a neighbourhood aggregation or message-passing scheme. Many variants based on the Kipf and Welling version of GCN have been proposed, but they can ultimately be expressed as neighbourhood aggregation or message passing schemes. For these different GCNs, I expect my hypothesis that a certain degree of alignment between ingredients is needed for them to perform well would hold since the working principles of variants of GCNs and the original version I consider are similar. To substantiate this claim, I have replicated my experiments using a recently proposed variant of GCN: SGC proposed by Ref. (Wu et al., 2019). SGC is a simplified version of the original GCN proposed by Kipf and Welling that removes nonlinearities and collapses weight matrices between consecutive layers. It has been shown that SGC can achieve competitive performance on node classification tasks and yields up to several orders of magnitude speedup.

I use the implementation provided by PyTorch Geometric[1], which is a popular GDL extension library for PyTorch. my results on SGC are shown below in Figure A.2 and Figure A.3. The figure suggests that results based on SGC are consistent with those produced using GCN (Kipf and Welling, 2017).

A.3 Choices of distance measures

Within the distances discussed by Ref. (Ye and Lim, 2016), there are two “families”:

1. average distances that use *all* the principal angles, e.g., the Chordal distance, $\left(\sum_{j=1}^{\alpha} \sin^2 \theta_j\right)^{1/2}$, and the Grassmann distance, $\left(\sum_{j=1}^{\alpha} \theta_j^2\right)^{1/2}$.

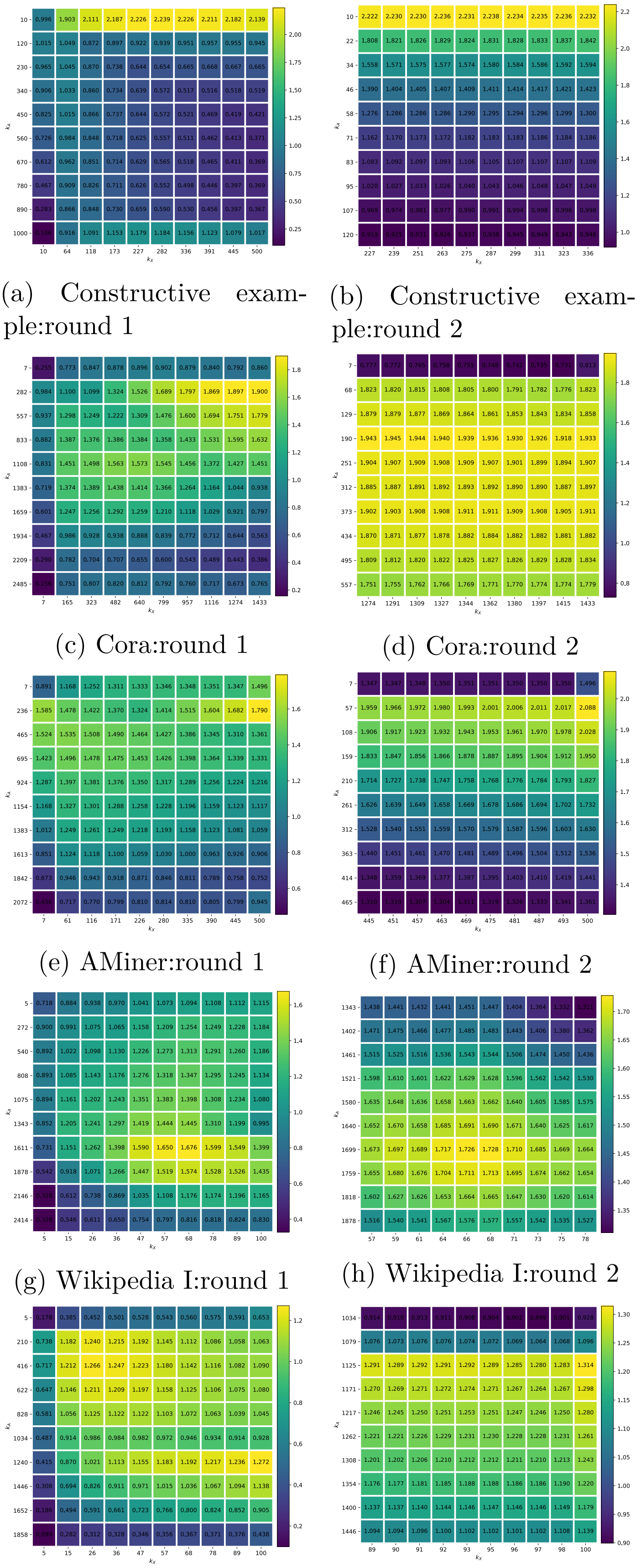
2. extremal distances that use only the maximum principal angle between two subspaces, e.g., the Projection distance, $\sin \theta_\alpha$where θ_α is the maximum angle.

[1] https://github.com/rusty1s/pytorch_geometric/blob/master/examples/sgc.py

my numerics show that average distances (the first family) display similar performance, as they leverage information from all the principal angles. Hence these measures produce a similar performance to the Chordal distance. To show this, I have replicated my experiments using the Grassmann distance (see Figure A.4 below). The results are consistent with those produced with Chordal distance.

On the other hand, I expect that extremal distances (the second family) will have less expressive power to capture the alignment between subspaces since they use solely the maximum principal angle and do not consider the information contained in the other principal angles. To demonstrate this point, I replicated my experiments with the Projection distance (see Figure A.5 below). my results show that the Projection distance is indeed less effective than the Chordal distance in representing the alignment between subspaces.

A.4 Supplemental tables and figures



(a) Constructive example:round 1

(b) Constructive example:round 2

(c) Cora:round 1

(d) Cora:round 2

(e) AMiner:round 1

(f) AMiner:round 2

(g) Wikipedia I:round 1

(h) Wikipedia I:round 2

(i) Wikipedia II:round 1

(j) Wikipedia II:round 2

Figure A.1: Summary of results on scanning subspaces.

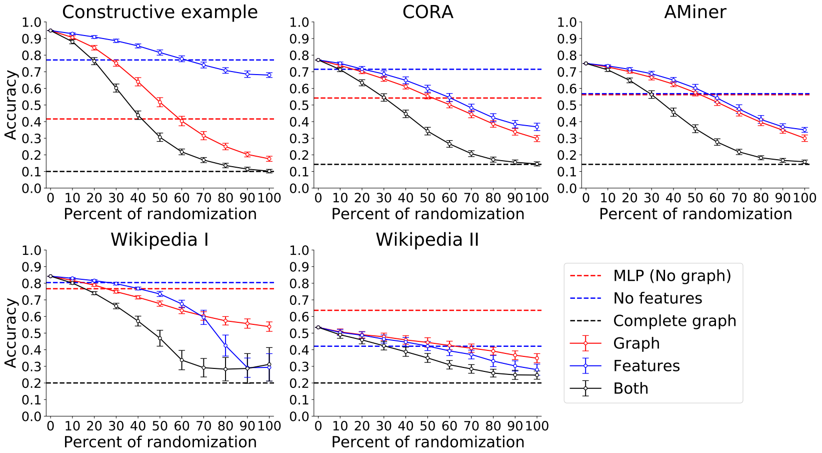


Figure A.2: **Degradation of the classification performance as a function of randomisation with SGC.** Each panel shows the degradation of the classification accuracy as a function of the randomisation of graph, features and both, for a different data set. Error bars are evaluated over 100 realisations: for zero percent randomisation, I report 100 runs with random seeds; for the rest, I report 1 run with random seed for 100 random realisations. The horizontal lines correspond to the limiting cases.

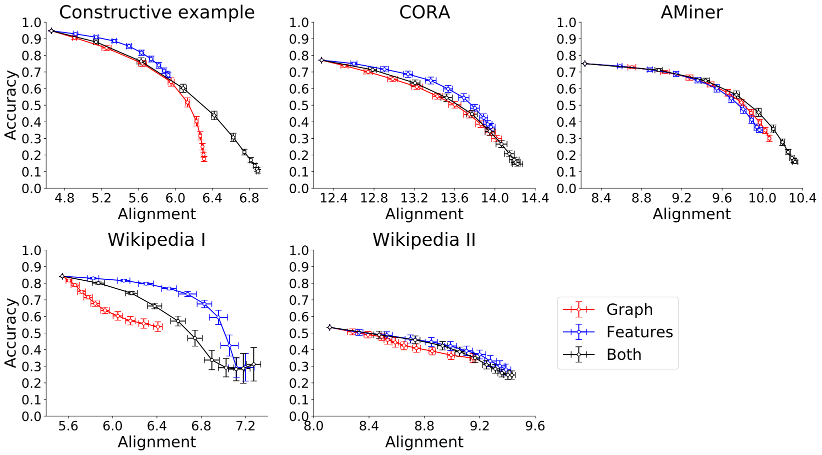


Figure A.3: **Classification performance versus the SAM with SGC.** Each panel shows the accuracy of SGC versus the SAM for all the runs presented in Figure A.2. Error bars are evaluated over 100 randomisations.

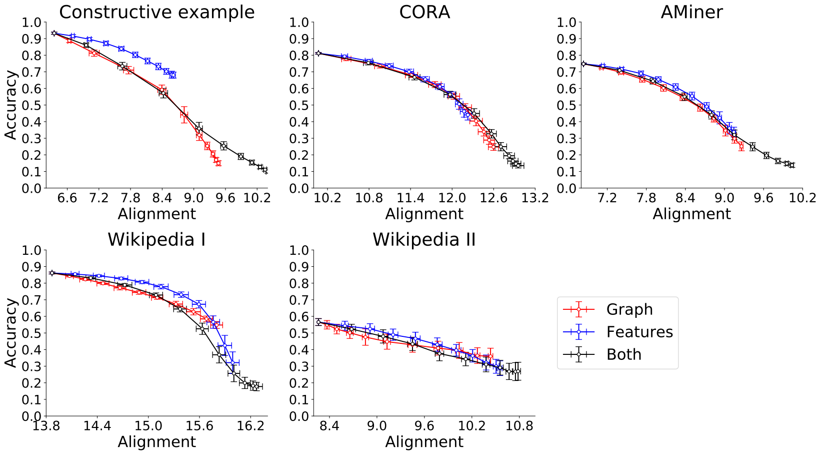


Figure A.4: **Classification performance versus the SAM with Grassmann distance.** Each panel shows the accuracy of GCN versus the SAM. Error bars are evaluated over 100 randomisations.

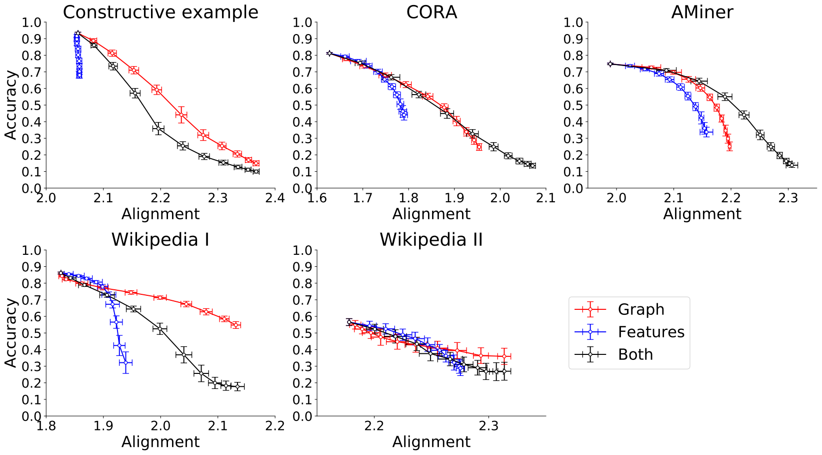


Figure A.5: **Classification performance versus the SAM with Projection distance.** Each panel shows the accuracy of GCN versus the SAM. Error bars are evaluated over 100 randomisations.

Appendix B

Appendix of Chapter 7

B.1 Comparison with Seurat clustering

Single-cell clustering is indeed an area where some of the gold standard methods are based on applying community detection to graphs derived from cell features (e.g., gene expression levels) by using the Louvain algorithm to maximise modularity. Seurat (Satija et al., 2015) is such a graph-based clustering approach, where a kNN graph is constructed from the PCA decomposition of the original feature vectors, and the obtained graph is then partitioned into communities (corresponding to cell types) by Louvain modularity maximisation.

It is important to remark that there is a fundamental distinction between the Seurat setting and my work. While my method (CkNN+GCN) addresses a classification problem (supervised setting, in which some class labels are known as ground truths and used in the training), Seurat solves a clustering problem (unsupervised setting, in which there are no class labels available on which to train).

Clustering aims to group similar samples together and dissimilar samples into distinct groups based on the similarity between their features (Liu and Barahona, 2020). Seurat clustering involves three steps: (i) compute the principal components of the feature vectors, and select the top T principal components based on a choice of p, the ratio of explained variance to total variance; (ii) construct a kNN graph based on the Euclidean distance between the vectors defined by the top T principal components of each sample; and (iii) perform community detection on the kNN graph using Louvain modularity maximisation. In this process, several hyperparameters are chosen, including the ratio p, which determines the number of principal components in step (i), and the number of neighbours k in the kNN graph in step (ii). The final result of Seurat is a partition of the data set into clusters ("graph communities") derived intrinsically from the properties of the data.

My method (CkNN+GCN), on the other hand, attempts a classification task where I leverage both the features and a feature-derived CkNN graph of appropriate edge density to train the weights of a GCN in order to maximise its classification power. My use of GCN and CkNN is distinctive in this setting, as is the optimisation of the edge density of the graph to maximise the quality of the classification. Given that the objectives of Seurat and my method are different, it is not straightforward to compare both approaches, but I have produced a setting to compare both methods. In particular, I have devised a comparison between my CkNN+GCN method and two Louvain-based clustering methods: Seurat (PCA+kNN+Louvain) and a simpler application of Louvain to a kNN graph of features (kNN+Louvain) without applying PCA in the first step. These three methods are compared to a simple kNN classifier (kNNC).

To compare the methods, I use the labels in the training and validation sets (defined as above in my CkNN+GCN experiments) as ground truths and tune the hyperparameters k and p (k is grid-searched over $[2, 4, 8, 16, 32, 64]$, and p is grid-searched over $[0.5, 0.6, 0.7, 0.8, 0.9]$) to maximise the similarity between the obtained clusters and the ground truth partitions. Once the hyperparameters have been optimised, I then use each method to cluster the data, and I compute the quality of the clustering against the test set. To evaluate the quality of the clustering, I use two standard measures: the Adjusted Rand Index (ARI) and the Normalised Mutual Information (NMI). Both of these measures are normalised between 0 (random assignment) and 1 (perfect agreement), with higher values signifying better assignment. My results are presented in Table B.5. My results show that my method (CkNN+GCN) performs better on average than both Louvain-based clustering methods on my data sets. However, CkNN+GCN does not always outperform the other methods; in particular, Seurat is the best on the Cell data set.

This might reflect particularities of the Cell data set, which contains high-dimensional vectors with high levels of noise that might benefit from the effective dimensionality reduction and filtering that PCA enforces. On the other hand, CkNN+GCN has been kept as a broad-purpose method, i.e., not optimised for a particular type of data. For instance, I use default values for some GCN hyperparameters (learning rate, number of hidden units, drop out rate) without optimising them on each data set. The aim is to provide robust outcomes across diverse data sets, as shown in Table B.5. Hence, there is room to potentially optimise my method (CkNN+GCN) specifically for single-cell genomics, but I feel this falls beyond the scope of my current work and will be investigated in future research.

Still, I would like to remark that clustering and classification algorithms are not directly comparable since they have different objectives and learning contexts. Nonetheless, I hope that my additional experiments provide some insight into the comparison.

B.2 Supplemental tables and figures

Table B.1: Classification accuracy (in percent) on the test set (average and standard deviation over 10 runs with random initialisations) for 7 data sets with 8 classifiers (four graph-less methods; GCN with four graph constructions).

Classifier	**Constructive**	**Cora**	**AMiner**	**Digits**	**FMA**	**Cell**	**Segmentation**
MLP = GCN (No graph)	42.1 ± 1.2	54.2 ± 1.7	54.4 ± 1.1	82.0 ± 1.3	34.3 ± 0.8	79.5 ± 3.0	72.0 ± 2.4
kNNC	31.4 ± 0.0	38.2 ± 0.0	28.0 ± 0.0	88.3 ± 0.0	30.6 ± 0.0	58.7 ± 0.0	68.8 ± 0.0
SVM	40.0 ± 0.0	55.9 ± 0.0	51.4 ± 0.0	87.7 ± 0.0	35.3 ± 0.0	81.5 ± 0.0	87.7 ± 0.0
RF	36.3 ± 1.0	56.1 ± 1.2	47.7 ± 1.5	83.0 ± 0.5	33.0 ± 0.9	88.0 ± 0.7	88.8 ± 0.7
GCN (kNN)	53.9 ± 0.9	66.4 ± 0.6	59.2 ± 1.3	92.0 ± 0.4	35.6 ± 1.0	83.8 ± 1.6	83.5 ± 0.7
GCN (MkNN)	45.2 ± 1.6	64.1 ± 0.4	61.8 ± 0.8	93.2 ± 0.3	35.6 ± 0.7	84.0 ± 2.0	83.0 ± 0.6
GCN (CkNN)	51.1 ± 1.3	66.6 ± 0.4	61.6 ± 0.8	93.4 ± 0.3	36.0 ± 0.8	84.0 ± 2.1	83.9 ± 0.6
GCN (RMST)	45.9 ± 1.5	64.8 ± 0.6	61.5 ± 1.3	89.3 ± 0.5	35.4 ± 0.7	84.9 ± 1.1	83.0 ± 1.6

Table B.2: Selected density parameters and density of constructed graphs in the graph densification process (Supplemental Figure B.1)

	kNN		**MkNN**		**CkNN** ($\delta = 1$)		**RMST** ($k = 1$)	
Data set	k^*	Density	k^*	Density	k^*	Density	γ^*	Density
Constructive	9	0.01741	104	0.02101	33	0.00920	0.07421	0.02724
Cora	12	0.00842	39	0.00436	74	0.01476	0.02924	0.01242
AMiner	8	0.00748	199	0.01786	199	0.03852	0.02317	0.00859
Digits	5	0.00404	39	0.01400	33	0.01564	0.00346	0.00117
FMA	1	0.00100	2	0.00103	13	0.00398	0.00146	0.00107
Cell	1	0.00100	8	0.00133	41	0.00753	0.00320	0.00124
Segmentation	7	0.00387	20	0.00637	12	0.00447	0.03423	0.00117

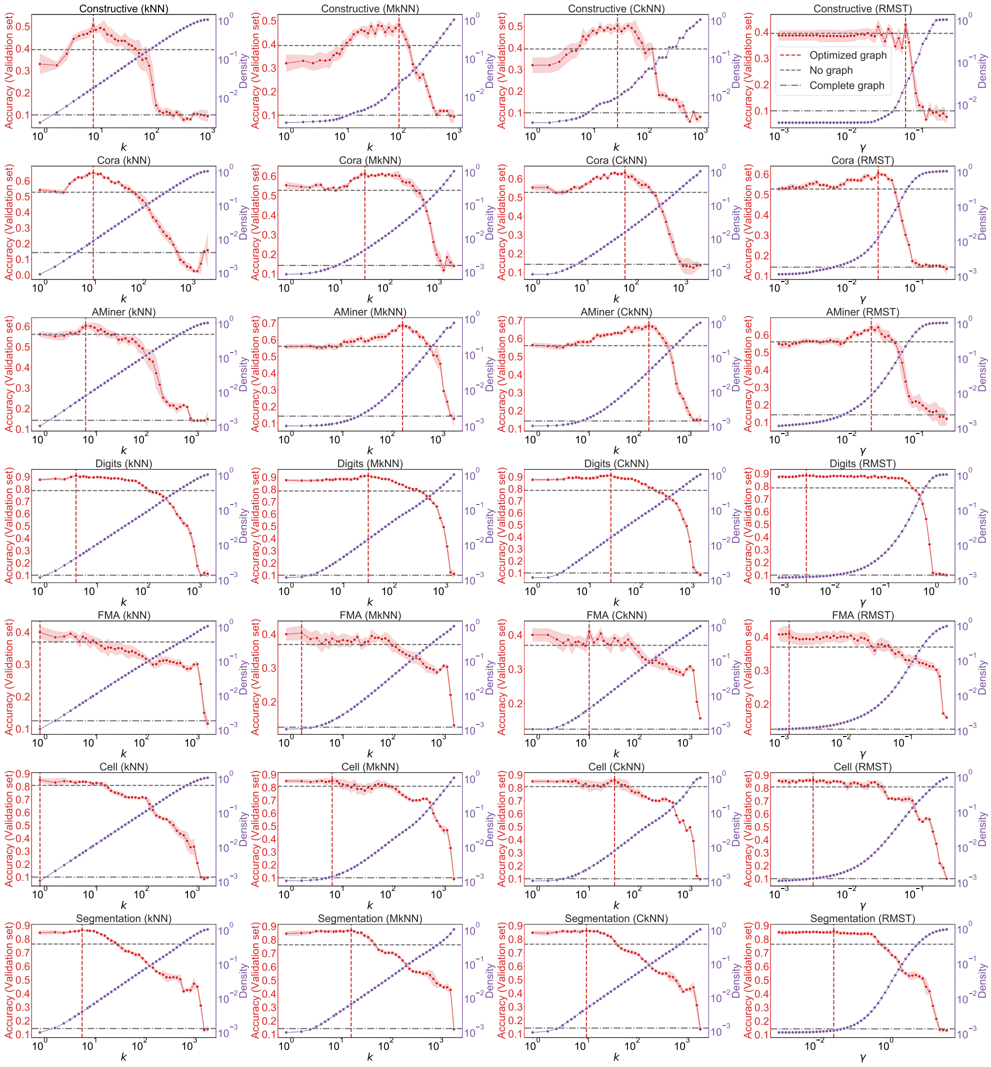


Figure B.1: Graph construction search in the densification process. The red line indicates the mean classification accuracy on the validation set of 10 runs with random weight initialisations as a function of the density parameter. The red shaded regions denote the standard deviation. The mean classification accuracy on the validation of two limiting cases (no graph and complete graph) is also added. The red vertical line indicates the optimised graph. The purple line shows the densities of the constructed graphs.

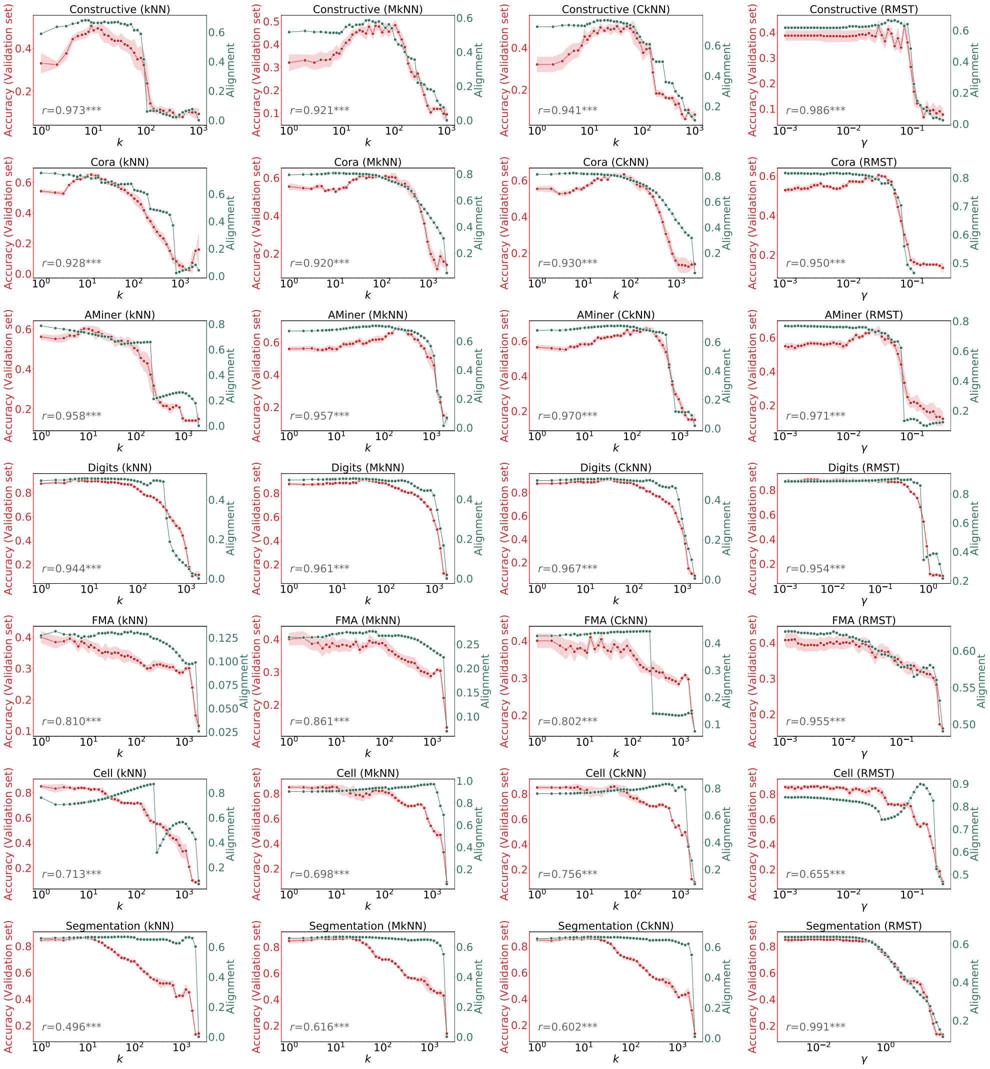


Figure B.2: The red line indicates the mean classification accuracy on the validation set of 10 runs with random weight initialisations as a function of the density parameter. The red shaded regions denote the standard deviation. The green line indicates the alignment. I report the Pearson correlation coefficients and p-values between mean accuracy and alignment. *p-value < 0.05,** p-value $<$ 0.01,*** p-value < 0.001.

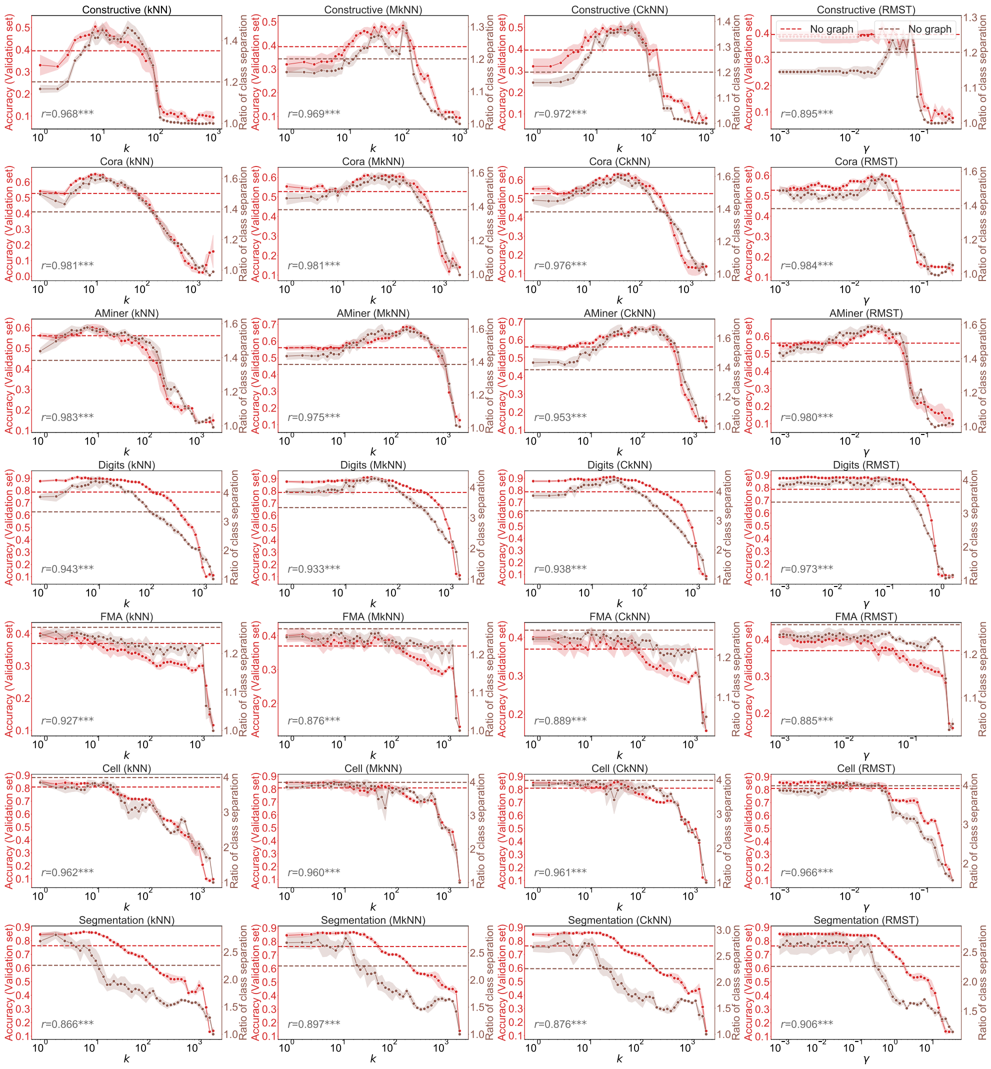


Figure B.3: The red line indicates the mean classification accuracy on the validation set of 10 runs with random weight initialisations as a function of the density parameter. The red shaded regions denote the standard deviation. The red dashed line represents the mean classification accuracy on the validation of no graph case. The brown line shows the ratio of class separation in the densification process. The brown shaded regions denote the standard deviation. The brown dashed line represents the ratio of class separation of no graph case. I report the Pearson correlation coefficients and p-values between mean accuracy and mean ratio of class separation. *p-value < 0.05,** p-value < 0.01,*** p-value < 0.001.

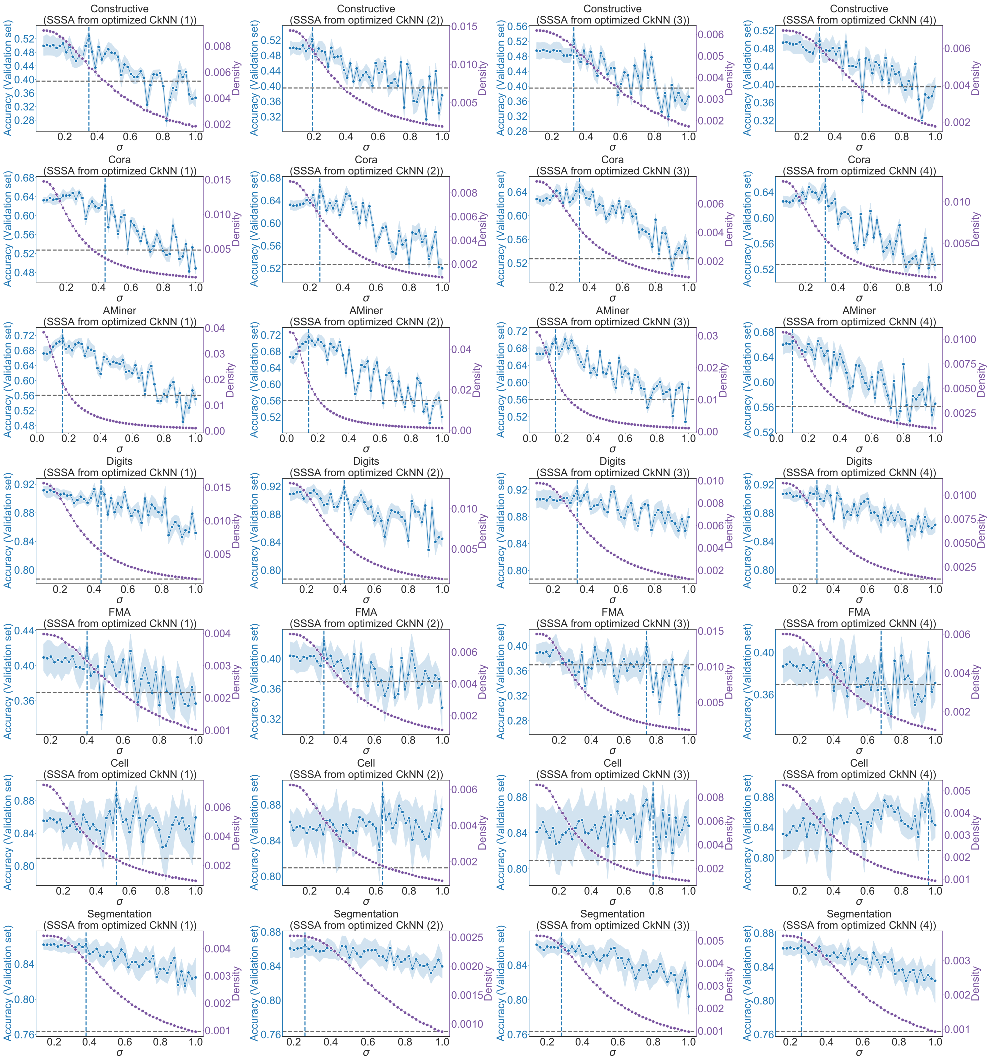


Figure B.4: Graph construction search in the sparsification process. The blue line indicates the mean classification accuracy on the validation set of 10 runs with random weight initialisations as a function of the density parameter. The blue shaded regions denote the standard deviation. The mean classification accuracy on the validation of no graph is added as well. The blue vertical line indicates the optimised graph on the validation set. The purple line shows the densities of the sparsified graphs.

Table B.3: Comparison between optimised CkNN and sparsification of optimised CkNN graphs (Supplemental Figure B.4).

Top 4 CkNN graphs on validation set	**Data set**	**Optimised CkNN** k^*	 Edge density	 ⟨Degree⟩	 Accuracy (Test)	**Sparsification of optimised CkNN** σ^*	 Edge density	 ⟨Degree⟩	 Accuracy (Test)
(1)	Constructive	33	0.00920	9.2	51.1	0.3479	0.00630	6.3	51.6
	Cora	74	0.01476	36.7	66.6	0	0.01476	36.7	66.6
	AMiner	199	0.03852	79.8	61.6	0.1618	0.01840	38.1	62.5
	Digits	33	0.01564	28.1	93.4	0	0.01564	28.1	93.4
	FMA	13	0.00398	8.0	36.0	0	0.00398	8.0	36.0
	Cell	41	0.00753	15.0	84.0	0.5212	0.00240	4.8	85.0
	Segmentation	12	0.00447	10.3	83.9	0.3806	0.00356	8.2	84.0
	Average improvement				(+8.3)				(+8.7)
(2)	Constructive	51	0.01445	14.4	51.8	0.1898	0.01197	12.0	53.6
	Cora	46	0.00897	22.3	66.3	0	0.00897	22.3	66.3
	AMiner	233	0.04838	100.2	61.3	0.1418	0.02396	49.6	63.6
	Digits	28	0.01319	23.7	93.2	0.4222	0.00556	10.0	93.2
	FMA	22	0.00713	14.3	35.2	0.3018	0.00561	11.2	35.8
	Cell	35	0.00625	12.5	83.6	0.6409	0.00176	3.5	86.9
	Segmentation	7	0.00253	5.8	84.0	0.2607	0.00252	5.8	84.2
	Average improvement				(+8.1)				(+9.3)
(3)	Constructive	16	0.00618	6.2	49.0	0	0.00618	6.2	49.0
	Cora	39	0.00756	18.8	66.8	0	0.00756	18.8	66.8
	AMiner	171	0.03115	64.5	62.1	0.1618	0.01656	34.3	63.5
	Digits	21	0.00978	17.6	92.9	0.3425	0.00650	11.7	93.0
	FMA	41	0.01457	29.1	35.9	0	0.01457	29.1	35.9
	Cell	48	0.00904	18.1	81.9	0.7806	0.00141	2.8	84.1
	Segmentation	14	0.00522	12.1	83.8	0	0.00522	12.1	83.8
	Average improvement				(+7.7)				(+8.2)
(4)	Constructive	22	0.00697	7.0	51.2	0.3084	0.00605	6.0	51.4
	Cora	63	0.01246	30.9	65.9	0	0.01246	30.9	65.9
	AMiner	78	0.01071	22.2	62.0	0.1019	0.01021	21.1	62.1
	Digits	24	0.01125	20.2	92.9	0	0.01125	20.2	92.9
	FMA	19	0.00601	12.0	34.5	0.6808	0.00201	4.0	35.2
	Cell	30	0.00527	10.5	81.8	0.9601	0.00099	2.0	85.3
	Segmentation	10	0.00372	8.6	83.9	0.2607	0.00365	8.4	84.1
	Average improvement				(+7.7)				(+8.3)

Table B.4: Classification accuracy (test set) obtained with three free feature-only methods: MLP, kNN+LDS+GCN (Franceschi et al., 2019), and CkNN+GCN (Chapter 7). For information, I also include the accuracy achieved by GCN applied to features together with the additional graph given in the original data set (when available).

Method	**Cora**	**AMiner**	**Digits**	**FMA**	**Cell**	**Segmentation**	**Avg. improvement**
MLP	54.2	54.4	82.0	34.3	79.5	72.0	—
kNN+LDS+GCN (Franceschi et al., 2019)	**69.0**	59.3	**94.6**	**36.2**	80.0	**83.9**	(+7.8)
CkNN + GCN (Chapter 7)	66.6	**61.6**	93.4	36.0	**84.0**	**83.9**	**(+8.2)**
Additional original graph + GCN	*81.1*	*74.8*	–	–	–	–	–

Table B.5: Quality of assignments (test set) obtained by a simple kNN classifier (kNNC), two Louvain-based methods (kNN+Louvain, Seurat), and my method (CkNN+GCN). The hyperparameters of all methods (kNNC, Louvain methods, and CkNN+GCN) were optimised on the training and validation sets. Two quality measures are computed (ARI and NMI), both normalised between 0 and 1, with higher values indicating better agreement with the ground truth of the test set. The average improvement with respect to the kNNC is also presented in the last column.

			ARI				
Method	**Cora**	**AMiner**	**Digits**	**FMA**	**Cell**	**Segmentation**	**Avg. improvement**
kNNC	0.090	0.036	0.766	0.087	0.434	0.456	—
kNN+Louvain	0.337	0.301	0.840	0.086	0.721	0.273	0.115
Seurat=PCA+kNN+Louvain	0.321	0.305	**0.888**	0.086	**0.822**	0.189	0.124
CkNN+GCN (Chapter 7)	**0.382**	**0.348**	0.863	**0.108**	0.767	**0.702**	**0.217**
			NMI				
Method	**Cora**	**AMiner**	**Digits**	**FMA**	**Cell**	**Segmentation**	**Avg. improvement**
kNNC	0.130	0.131	0.806	0.123	0.644	0.532	—
kNN+Louvain	0.386	0.323	0.892	0.125	0.811	0.513	0.114
Seurat=PCA+kNN+Louvain	0.391	0.356	**0.904**	0.118	**0.904**	0.350	0.110
CkNN+GCN (Chapter 7)	**0.408**	**0.409**	0.889	**0.147**	0.854	**0.753**	**0.183**

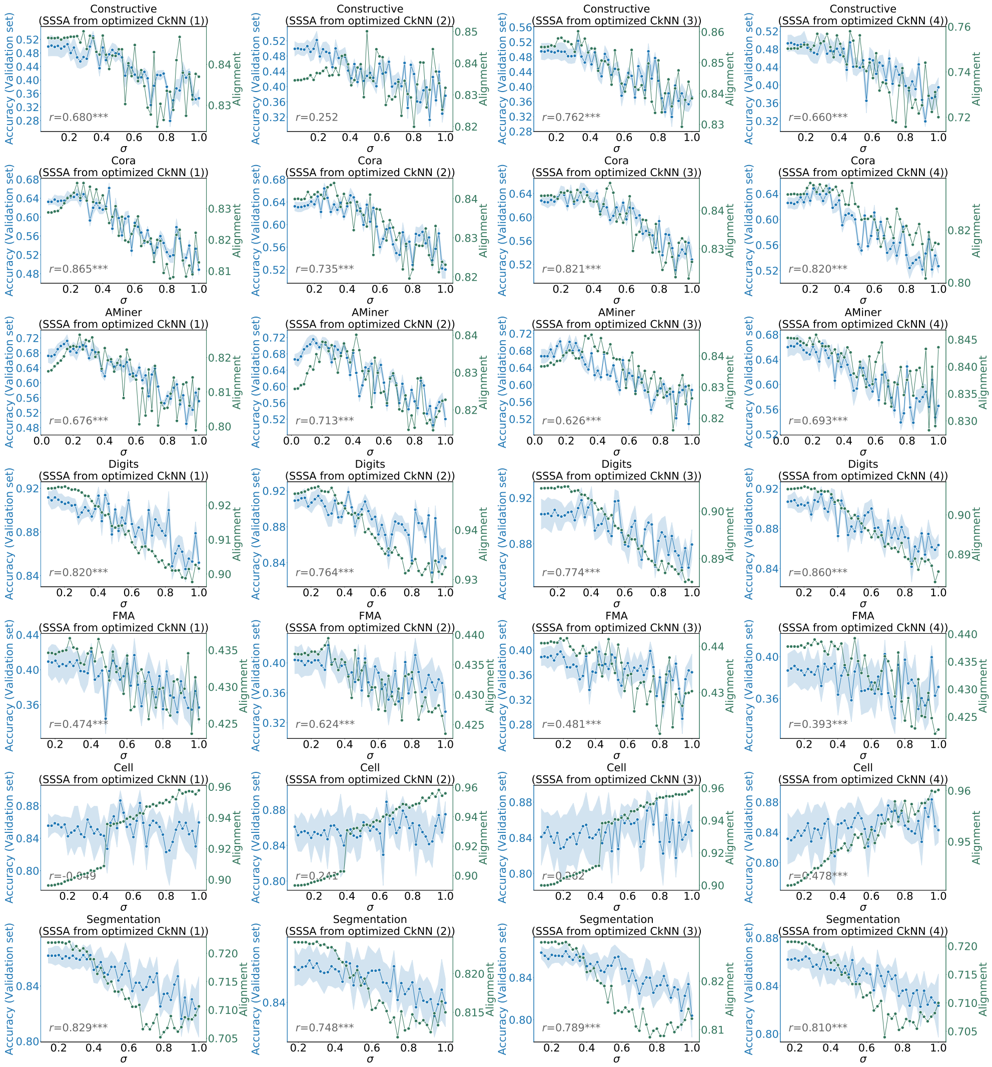


Figure B.5: The blue line indicates the mean classification accuracy on the validation set of 10 runs with random weight initialisations as a function of the density parameter. The blue shaded regions denote the standard deviation. The green line indicates the alignment. I report the Pearson correlation coefficients and p-values between mean accuracy and alignment. *p-value < 0.05,** p-value < 0.01,*** p-value < 0.001.

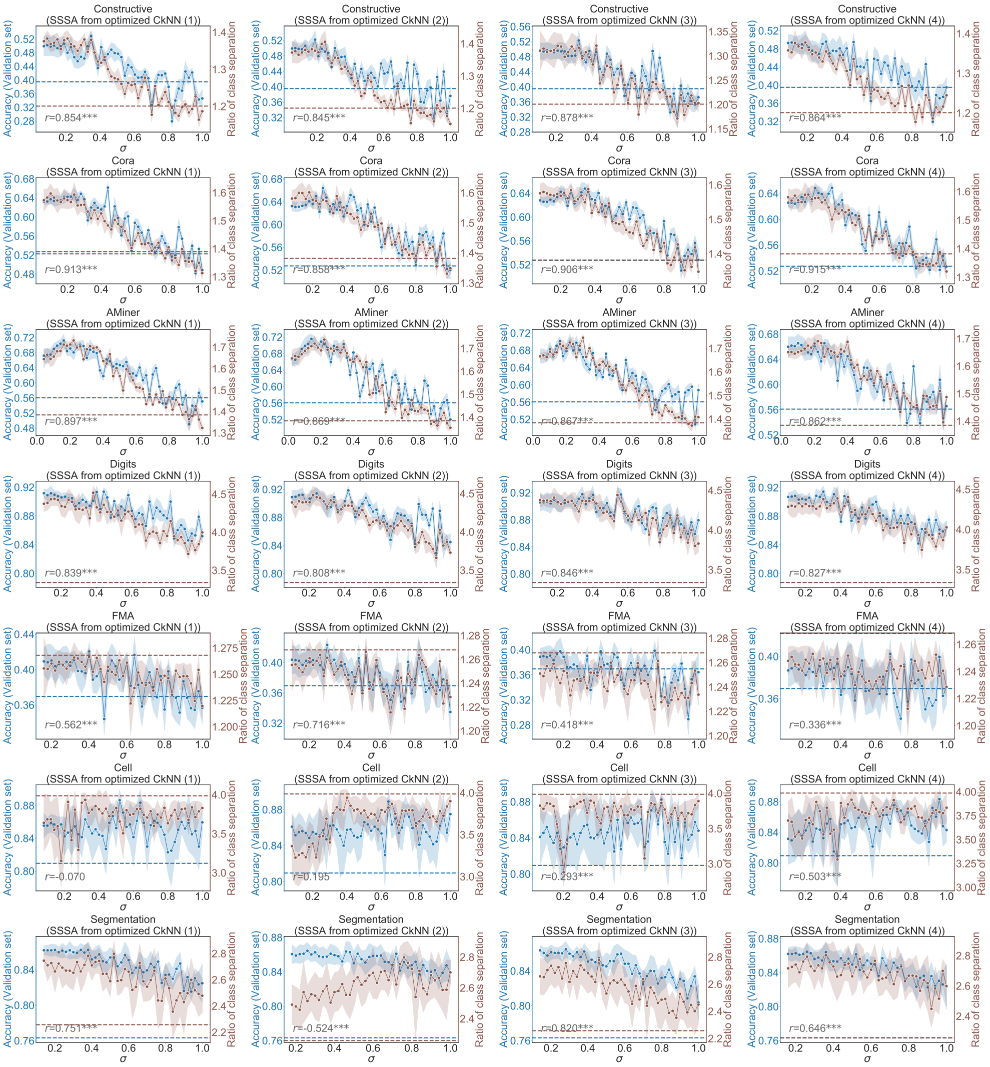


Figure B.6: The blue line indicates the mean classification accuracy on the validation set of 10 runs with random weight initialisations as a function of the density parameter. The blue shaded regions denote the standard deviation. The blue dashed line represents the mean classification accuracy on the validation of no graph case. The brown line shows the ratio of class separation in the sparsification process. The brown shaded regions denote the standard deviation. The brown dashed line represents the ratio of class separation of no graph case. I report the Pearson correlation coefficients and p-values between mean accuracy and mean ratio of class separation. *p-value < 0.05,** p-value < 0.01,*** p-value < 0.001.

Appendix C

Publications

Journal publications

- **Yifan Qian**, Paul Expert, Tom Rieu, Pietro Panzarasa, and Mauricio Barahona. Quantifying the alignment of graph and features in deep learning. ***IEEE Transactions on Neural Networks and Learning Systems*** *(2021).* doi: https://doi.org/10.1109/TNNLS.2020.3043196

- **Yifan Qian**, Paul Expert, Pietro Panzarasa, and Mauricio Barahona. Geometric graphs from data to aid classification tasks with graph convolutional networks. ***Patterns Cell Press*** 2, no. 4 (2021): 100237. doi: https://doi.org/10.1016/j.patter.2021.100237

Working papers

- **Yifan Qian**, Luca Verginer, Xiancheng Li, Massimo Riccaboni, and Pietro Panzarasa. Network foundations of scientific impact and innovation of cities. *In Preparation for Submission.*

- **Yifan Qian** and Pietro Panzarasa. Intra- and inter-brokerage in social networks. *In Preparation for Submission.*

- **Yifan Qian**, Marco Serino, Leslie DeChurch, Noshir Contractor, and Pietro Panzarasa. The structural foundations of creativity: A network-based study of co-productions among Italian theatres. *Data Analysis.*

- **Yifan Qian**, Luca Verginer, Greg Morrison, Massimo Riccaboni, and Pietro Panzarasa. Social distance with editors in co-authorship networks. *Data Analysis.*

- **Yifan Qian**, Ching Jin, and Pietro Panzarasa. Teams with unexpected collaboration between institutions are associated with higher impact. *Data Analysis.*

- Xiancheng Li, **Yifan Qian**, Mauricio Barahona, and Pietro Panzarasa. Publishing the first patent boosts scientists' performance. *Data Analysis.*

- Hongwei Peng, **Yifan Qian**, and Pietro Panzarasa. Quantifying the temporal patterns of interdisciplinarity in scientists' careers. *Data Analysis.*

Appendix D

Conference presentations

- Chapter 3 was presented in: *IC2S2 (2019)*, Poster.
- Chapter 4 was presented in: *NetSci (2021)*, Talk; *NetSci (2020)*, Poster; *IC2S2 (2020)*, Poster; *IC2S2 (2019)*, Talk.
- Chapter 6 was presented in: *IC2S2 (2019)*, Poster; *Data Natives (2019)*, Talk; *UK Network Science workshop (2018)*, Talk; *IC2S2 (2018)*, Poster; *NetSci Satellite on Machine Learning in Network Science (2018)*, Talk; *NetSci (2018)*, Poster.
- Chapter 7 was presented in: *NetSci (2020)*, Talk; *IC2S2 (2020)*, Poster.

NetSci: International School and Conference on Network Science.

IC2S2: International Conference on Computational Social Science.

Bibliography

Abbasi, A. and Jaafari, A. (2013). Research impact and scholars' geographical diversity. *Journal of Informetrics*, 7(3):683–692.

Adams, J. (2013). Collaborations: The fourth age of research. *Nature*, 497(7451):557.

Adams, J. D., Black, G. C., Clemmons, J. R., and Stephan, P. E. (2005). Scientific teams and institutional collaborations: Evidence from us universities, 1981–1999. *Research Policy*, 34(3):259–285.

Allamanis, M., Brockschmidt, M., and Khademi, M. (2018). Learning to represent programs with graphs. In *International Conference on Learning Representations (ICLR)*.

AlShebli, B. K., Rahwan, T., and Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature Communications*, 9(1):1–10.

Andoni, A. and Indyk, P. (2006). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *47th annual IEEE symposium on foundations of computer science*, pages 459–468. IEEE.

Aral, S. and Van Alstyne, M. (2011). The diversity-bandwidth trade-off. *American Journal of Sociology*, 117(1):90–171.

Atwood, J. and Towsley, D. (2016). Diffusion-convolutional neural networks. In *Neural Information Processing Systems (NeurIPS)*, pages 1993–2001.

Barjak, F. and Robinson, S. (2008). International collaboration, mobility and team diversity in the life sciences: impact on research performance. *Social Geography*, 3(1):23–36.

Batjargal, B. (2007). Internet entrepreneurship: Social capital, human capital, and performance of internet ventures in china. *Research policy*, 36(5):605–618.

Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G., and Petri, G. (2020). Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 874:1–92.

Baum, J. A., McEvily, B., and Rowley, T. J. (2012). Better with age? tie longevity and the performance implications of bridging and closure. *Organization Science*, 23(2):529–546.

Beguerisse-Díaz, M., Vangelov, B., and Barahona, M. (2013). Finding role communities in directed networks using role-based similarity, markov stability and the relaxed minimum spanning tree. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 937–940. IEEE.

Berry, T. and Sauer, T. (2019). Consistent manifold representation for topological data analysis. *Foundations of Data Science*, 1(1):1–38.

Bettencourt, L. M. (2013). The origins of scaling in cities. *Science*, 340(6139):1438–1441.

Bettencourt, L. M., Lobo, J., Helbing, D., Kühnert, C., and West, G. B. (2007a). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306.

Bettencourt, L. M., Lobo, J., and Strumsky, D. (2007b). Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research Policy*, 36(1):107–120.

Björck, A. and Golub, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bollobás, B. (1998). *Modern Graph Theory*. Graduate texts in mathematics. Springer, New York, USA.

Borgatti, S. P. (1997). Structural holes: Unpacking burt's redundancy measures. *Connections*, 20(1):35–38.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.

Bruna, J. and Mallat, S. (2013). Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2014). Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR)*.

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, 110(2):349–399.

Burt, R. S. (2005). *Brokerage and Closure: An Introduction to Social Capital*. Oxford University Press, Oxford, England, UK.

Burt, R. S. (2009). *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, Massachusetts, USA.

Burt, R. S. (2010). *Neighbor Networks: Competitive Advantage Local and Personal*. Oxford University Press, Oxford, UK.

Burt, R. S. and Knez, M. (1995). Kinds of third-party effects on trust. *Rationality and Society*, 7(3):255–292.

Cantner, U. and Rake, B. (2014). International research networks in pharmaceuticals: Structure and dynamics. *Research Policy*, 43(2):333–348.

Catts, R. and Ozga, J. (2005). *What is Social Capital and How Might It Be Used in Scotland's Schools?*, volume 36. Centre for Educational Sociology Edinburgh, Leeds, UK.

Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., and Sun, X. (2020). Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Chen, J., Fang, H.-r., and Saad, Y. (2009). Fast approximate knn graph construction for high dimensional data via recursive lanczos bisection. *Journal of Machine Learning Research*, 10(9).

Chen, K., Zhang, Y., and Fu, X. (2019). International research collaboration: An emerging domain of innovation studies? *Research Policy*, 48(1):149–168.

Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. (2017). Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795.

Choi, E., Xiao, C., Stewart, W., and Sun, J. (2018). Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Neural Information Processing Systems (NeurIPS)*, pages 4547–4557.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, pages S95–S120.

Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press.

Cummings, J. N. (2004). Work groups, structural diversity, and knowledge sharing in a global organization. *Management Science*, 50(3):352–364.

Davis, J. A. (1970). Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices. *American Sociological Review*, pages 843–851.

Davis, J. A., Holland, P., and Leinhardt, S. (1971). Comments on professor mazur's hypothesis about interpersonal sentiments. *American Sociological Review*, 36(2):309–311.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2017). Fma: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Neural Information Processing Systems (NeurIPS)*, pages 3844–3852.

Degenne, A. and Forsé, M. (1999). *Introducing social networks*. Sage.

Delvenne, J.-C., Yaliraki, S. N., and Barahona, M. (2010). Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 107(29):12755–12760.

Deng, L. and Yu, D. (2014). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, 7(3-4):197–387.

Donetti, L., Hurtado, P. I., and Munoz, M. A. (2005). Entangled networks, synchronization, and optimal network topology. *Physical Review Letters*, 95(18):188701.

Drucker, P. F. (1994). *Post-Capitalist Society*. HarperCollins Publishers, New York, NY, USA.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *Neural Information Processing Systems (NeurIPS)*, pages 2224–2232.

Edler, J., Fier, H., and Grimpe, C. (2011). International scientist mobility and the locus of knowledge and technology transfer. *Research Policy*, 40(6):791–805.

Eisenhardt, K. M. and Tabrizi, B. N. (1995). Accelerating adaptive processes: Product innovation in the global computer industry. *Administrative Science Quarterly*, pages 84–110.

Evans, T., Lambiotte, R., and Panzarasa, P. (2011). Community structure and patterns of scientific collaboration in business and management. *Scientometrics*, 89(1):381–396.

Fainstein, S. S. (2005). Cities and diversity: should we want it? can we plan for it? *Urban affairs review*, 41(1):3–19.

Field, D. J. (1989). What the statistics of natural images tell us about visual coding. In *Human Vision, Visual Processing, and Digital Display*, volume 1077, pages 269–277.

Fleming, L., Mingo, S., and Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly*, 52(3):443–475.

Florida, R. (2002). *The rise of the creative class*, volume 9. Basic Books New York.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, 359(6379).

Franceschi, L., Niepert, M., Pontil, M., and He, X. (2019). Learning discrete structures for graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Frank, M. R., Wang, D., Cebrian, M., and Rahwan, I. (2019). The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence*, 1(2):79–85.

Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science*, 1.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

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

Fung, W.-S., Hariharan, R., Harvey, N. J., and Panigrahi, D. (2019). A general framework for graph sparsification. *SIAM Journal on Computing*, 48(4):1196–1223.

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M., and Correia, B. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192.

Gamst, F. C. (1991). Foundations of social theory. *Anthropology of Work Review*, 12(3):19–25.

Gargiulo, M. and Benassi, M. (2000). Trapped in your own net? network cohesion, structural holes, and the adaptation of social capital. *Organization Science*, 11(2):183–196.

Gargiulo, M., Ertug, G., and Galunic, C. (2009). The two faces of control: Network closure and individual performance among knowledge workers. *Administrative Science Quarterly*, 54(2):299–333.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*, volume 3. JHU Press.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gori, M., Monfardini, G., and Scarselli, F. (2005). A new model for learning in graph domains. In *Proceedings of IEEE International Joint Conference on Neural Networks*, volume 2, pages 729–734. IEEE.

Gould, R. V. and Fernandez, R. M. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, pages 89–126.

Graf, H. and Kalthaus, M. (2018). International research networks: Determinants of country embeddedness. *Research Policy*, 47(7):1198–1214.

Granovetter, M. (1977). The strength of weak ties. In *Social Networks*, pages 347–367. Elsevier.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3):481–510.

Granovetter, M. (2005). The impact of social structure on economic outcomes. *Journal of Economic Perspectives*, 19(1):33–50.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

Guan, J. and Liu, N. (2016). Exploitative and exploratory innovations in knowledge network and collaboration network: A patent analysis in the technological field of nano-energy. *Research policy*, 45(1):97–112.

Guan, J., Zhang, J., and Yan, Y. (2015). The impact of multilevel networks on innovation. *Research Policy*, 44(3):545–559.

Guan, J., Zuo, K., Chen, K., and Yam, R. C. (2016). Does country-level r&d efficiency benefit from the collaboration network structure? *Research Policy*, 45(4):770–784.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. In *Neural Information Processing Systems (NeurIPS)*, pages 1024–1034.

Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150.

Hansen, M. T. (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly*, 44(1):82–111.

Harris, R. G. (2001). The knowledge-based economy: Intellectual origins and new economic perspectives. *International Journal of Management Reviews*, 3(1):21–40.

Holland, P. W. and Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies*, 2(2):107–124.

Holland, P. W. and Leinhardt, S. (1977). A method for detecting structure in sociometric data. In *Social Networks*, pages 411–432. Elsevier.

Hoory, S., Linial, N., and Wigderson, A. (2006). Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561.

Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P., and Mascolo, C. (2016). Measuring urban social diversity using interconnected geo-social networks. In *Proceedings of the 25th International Conference on World Wide Web*, pages 21–30.

Hur, W. and Oh, J. (2021). A man is known by the company he keeps?: A structural relationship between backward citation and forward citation of patents. *Research Policy*, 50(1):104117.

Ingram, P. and Roberts, P. W. (2000). Friendships among competitors in the sydney hotel industry. *American Journal of Sociology*, 106(2):387–423.

Jacobs, J. (1985). *Cities and the wealth of nations: Principles of economic life.* Vintage.

Jacobs, J. (2016). *The economy of cities.* Vintage.

Jia, T., Wang, D., and Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, 1(4):1–7.

Ke, Q., Liang, L., Ding, Y., David, S. V., and Acuna, D. E. (2021). A dataset of mentorship in science with semantic and demographic estimations. *arXiv preprint arXiv:2106.06487.*

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Kilduff, M. and Brass, D. J. (2010). Organizational social network research: Core ideas and key debates. *Academy of Management Annals*, 4(1):317–357.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR).*

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR).*

Knyazev, A. V. and Argentati, M. E. (2002). Principal angles between subspaces in an a-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040.

Kwon, S.-W. and Adler, P. S. (2014). Social capital: Maturation of a field of research. *Academy of Management Review*, 39(4):412–422.

Lambiotte, R., Delvenne, J.-C., and Barahona, M. (2014). Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76–90.

Lambiotte, R. and Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, 3(3):180–190.

Lambiotte, R., Rosvall, M., and Scholtes, I. (2019). From networks to optimal higher-order models of complex systems. *Nature physics*, 15(4):313–320.

Landrieu, L. and Simonovsky, M. (2018). Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4558–4567.

Latora, V., Nicosia, V., and Panzarasa, P. (2013). Social cohesion, structural holes, and a tale of two measures. *Journal of Statistical Physics*, 151(3-4):745–764.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Computational social science. *Science*, 323(5915):721–723.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Neural Information Processing Systems (NeurIPS)*, pages 396–404.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 177–187.

Leydesdorff, L. and Wagner, C. S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics*, 2(4):317–325.

Li, E. Y., Liao, C. H., and Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9):1515–1530.

Li, Q., Han, Z., and Wu, X.-M. (2018a). Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Li, R., Dong, L., Zhang, J., Wang, X., Wang, W.-X., Di, Z., and Stanley, H. E. (2017). Simple spatial scaling rules behind complex cities. *Nature Communications*, 8(1):1–7.

Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. S. (2016). Gated graph sequence neural networks. In *International Conference on Learning Representations (ICLR)*.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2018b). Diffusion convolutional recurrent neural network: data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR)*.

Li, Z., Chen, Q., and Koltun, V. (2018c). Combinatorial optimization with graph convolutional networks and guided tree search. In *Neural Information Processing Systems (NeurIPS)*, pages 539–548.

Liang, X. and Liu, A. M. (2018). The evolution of government sponsored collaboration network and its impact on innovation: A bibliometric analysis in the chinese solar pv sector. *Research Policy*, 47(7):1295–1308.

Lin, N. (2002). *Social Capital: A Theory of Social Structure and Action*, volume 19. Cambridge university press, Cambridge, UK.

Lingo, E. L. and O'Mahony, S. (2010). Nexus work: Brokerage on creative projects. *Administrative Science Quarterly*, 55(1):47–81.

Liu, Z. and Barahona, M. (2020). Graph-based data clustering via multiscale community detection. *Applied Network Science*, 5(1):3.

Lord, L.-D., Allen, P., Expert, P., Howes, O., Broome, M., Lambiotte, R., Fusar-Poli, P., Valli, I., McGuire, P., and Turkheimer, F. E. (2012). Functional brain networks before the onset of psychosis: a prospective fmri study with graph theoretical analysis. *NeuroImage: Clinical*, 1(1):91–98.

Lu, X. and McInerney, P.-B. (2016). Is it better to "stand on two boats" or "sit on the chinese lap"?: Examining the cultural contingency of network structures in the contemporary chinese academic labor market. *Research Policy*, 45(10):2125–2137.

Luce, R. D. and Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

McCabe, S., Torres, L., LaRock, T., Haque, S. A., Yang, C.-H., Hartle, H., and Klein, B. (2021). netrd: A library for network reconstruction and graph distances. *Journal of Open Source Software*, 6(62):2990.

McFadyen, M. A. and Cannella Jr, A. A. (2004). Social capital and knowledge creation: Diminishing returns of the number and strength of exchange relationships. *Academy of Management Journal*, 47(5):735–746.

McFadyen, M. A., Semadeni, M., and Cannella Jr, A. A. (2009). Value of strong ties to disconnected others: Examining knowledge creation in biomedicine. *Organization Science*, 20(3):552–564.

Monti, F., Bronstein, M., and Bresson, X. (2017). Geometric matrix completion with recurrent multi-graph neural networks. In *Neural Information Processing Systems (NeurIPS)*, pages 3697–3707.

Muller, E. and Zenker, A. (2001). Business services as actors of knowledge transformation: the role of kibs in regional and national innovation systems. *Research Policy*, 30(9):1501–1516.

Nahapiet, J. and Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *The Academy of Management Review*, 23(2):242–266.

Neal, Z. (2011). Differentiating centrality and power in the world city network. *Urban Studies*, 48(13):2733–2748.

Newman, M. (2018a). Network structure from rich but noisy data. *Nature Physics*, 14(6):542–545.

Newman, M. (2018b). *Networks*. Oxford University Press.

Newman, M. E. (2001a). Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 64(1):016131.

Newman, M. E. (2001b). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132.

Newman, M. E. (2001c). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132.

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

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Ni, C., Smith, E., Yuan, H., Larivière, V., and Sugimoto, C. R. (2021). The gendered nature of authorship. *Science advances*, 7(36):eabe4639.

Niepert, M., Ahmed, M., and Kutzkov, K. (2016). Learning convolutional neural networks for graphs. In *International Conference on Machine Learning (ICML)*, pages 2014–2023.

Obstfeld, D. (2005). Social networks, the tertius iungens orientation, and involvement in innovation. *Administrative Science Quarterly*, 50(1):100–130.

Omranian, N., Eloundou-Mbebi, J. M., Mueller-Roeber, B., and Nikoloski, Z. (2016). Gene regulatory network inference using fused lasso on multiple data sets. *Scientific Reports*, 6:20533.

Østergaard, C. R., Timmermans, B., and Kristinsson, K. (2011). Does a different view create something new? the effect of employee diversity on innovation. *Research Policy*, 40(3):500–509.

Ozaki, K., Shimbo, M., Komachi, M., and Matsumoto, Y. (2011). Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 154–162.

Pain, E. (2018). Collaborating for the win. *Science*.

Patania, A., Petri, G., and Vaccarino, F. (2017). The shape of collaborations. *EPJ Data Science*, 6:1–16.

Peach, R. L., Arnaudon, A., and Barahona, M. (2020). Semi-supervised classification on graphs using explicit diffusion dynamics. *Foundations of Data Science*, 2(1):19.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Perraudin, N., Paratte, J., Shuman, D., Martin, L., Kalofolias, V., Vandergheynst, P., and Hammond, D. K. (2014). Gspbox: A toolbox for signal processing on graphs. *arXiv preprint arXiv:1408.5781*.

Perry-Smith, J. E. and Shalley, C. E. (2003). The social side of creativity: A static and dynamic social network perspective. *Academy of management review*, 28(1):89–106.

Qian, Y., Expert, P., Panzarasa, P., and Barahona, M. (2021a). Geometric graphs from data to aid classification tasks with graph convolutional networks. *Patterns*, 2(4):100237.

Qian, Y., Expert, P., Rieu, T., Panzarasa, P., and Barahona, M. (2021b). Quantifying the alignment of graph and features in deep learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Qian, Y., Rong, W., Jiang, N., Tang, J., and Xiong, Z. (2017). Citation regression analysis of computer science publications in different ranking categories and subfields. *Scientometrics*, 110(3):1351–1374.

Qiu, J., Tang, J., Ma, H., Dong, Y., Wang, K., and Tang, J. (2018). Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2110–2119.

Radovanović, M., Nanopoulos, A., and Ivanović, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Reagans, R. and McEvily, B. (2003). Network structure and knowledge transfer:

The effects of cohesion and range. *Administrative Science Quarterly*, 48(2):240–267.

Roethlisberger, F. J. and Dickson, W. J. (1939). Management and the worker cambridge. *Harvard University.*

Salnikov, V., Cassese, D., Lambiotte, R., and Jones, N. S. (2018). Co-occurrence simplicial complexes in mathematics: identifying the holes of knowledge. *Applied network science*, 3(1):1–23.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502.

Scellato, G., Franzoni, C., and Stephan, P. (2015). Migrant scientists and international networks. *Research Policy*, 44(1):108–120.

Schilling, M. A. and Fang, C. (2014). When hubs forget, lie, and play favorites: Interpersonal network structure, information distortion, and organizational learning. *Strategic Management Journal*, 35(7):974–994.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance components*, volume 391. John Wiley & Sons.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. (2008). Collective classification in network data. *AI Magazine*, 29(3):93–93.

Shipilov, A. V. and Li, S. X. (2008). Can you have your cake and eat it too? structural holes' influence on status accumulation and market performance in collaborative networks. *Administrative Science Quarterly*, 53(1):73–108.

Simmel, G. (1950). *The sociology of georg simmel*, volume 92892. Simon and Schuster, New York City, USA.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216.

Singh, J. (2008). Distributed r&d, cross-regional knowledge integration and quality of innovative output. *Research Policy*, 37(1):77–96.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., and Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246.

Sosa, M. E. (2011). Where do creative interactions come from? the role of tie content and social networks. *Organization Science*, 22(1):1–21.

Spielman, D. A. and Srivastava, N. (2011). Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926.

Spielman, D. A. and Teng, S.-H. (2011). Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025.

Stovel, K. and Shaw, L. (2012). Brokerage. *Annual Review of Sociology*, 38:139–158.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998. ACM.

Ter Wal, A. L., Alexy, O., Block, J., and Sandner, P. G. (2016). The best of both worlds: The benefits of open-specialized and closed-diverse syndication networks for new ventures' success. *Administrative Science Quarterly*, 61(3):393–432.

Torvik, V. I. (2015). Mapaffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. In *D-Lib magazine: the magazine of the Digital Library Forum*, volume 21. NIH Public Access.

Torvik, V. I. and Smalheiser, N. R. (2009). Author name disambiguation in medline. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3):11.

Uzzi, B. (1996). The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American Sociological Review*, pages 674–698.

Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, pages 35–67.

Uzzi, B. and Spiro, J. (2005). Collaboration and creativity: The small world problem. *American Journal of Sociology*, 111(2):447–504.

Vaccario, G., Verginer, L., and Schweitzer, F. (2020). The mobility network of scientists: Analyzing temporal correlations in scientific careers. *Applied Network Science*, 5(1):1–14.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations (ICLR)*.

Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. (2019). Deep graph infomax. In *International Conference on Learning Representations (ICLR)*.

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H., and Kriegstein, A. R. (2019). Single-cell genomics identifies cell type–specific molecular changes in autism. *Science*, 364(6441):685–689.

Verginer, L. and Riccaboni, M. (2020a). Cities and countries in the global scientist mobility network. *Applied Network Science*, 5(1):1–16.

Verginer, L. and Riccaboni, M. (2020b). Talent goes to global cities: The world network of scientists' mobility. *Research Policy*, 50(1):104127.

Vespignani, A. (2018). Twenty years of network science.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wagner, C. S. and Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10):1608–1618.

Wang, J. (2016). Knowledge creation in collaboration networks: Effects of tie configuration. *Research Policy*, 45(1):68–80.

Wang, Y., Jones, B. F., and Wang, D. (2019a). Early-career setback and future career impact. *Nature Communications*, 10(1):1–10.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. (2019b). Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5):1–12.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, Cambridge, UK.

Watts, D. J. (1999). *Small Worlds: The Dynamics of Networks Between Order and Randomness.* Princeton University Press, Princeton, New Jersey, USA.

Watts, D. J. (2004). The “new” science of networks. *Annual Review of Sociology*, 30:243–270.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of ‘small-world’networks. *Nature*, 393(6684):440–442.

Wu, F., Zhang, T., Souza Jr, A. H. d., Fifty, C., Yu, T., and Weinberger, K. Q. (2019). Simplifying graph convolutional networks. In *International Conference on Machine Learning (ICML)*, pages 6861–6871.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems.*

Xu, D., Zhu, Y., Choy, C. B., and Fei-Fei, L. (2017). Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*.

Ye, K. and Lim, L.-H. (2016). Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197.

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). Gnn explainer: A tool for post-hoc explanation of graph neural networks. *arXiv preprint arXiv:1903.03894*.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983.

Young, I. M. (2011). *Justice and the Politics of Difference*. Princeton University Press.

Yu, B., Yin, H., and Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*.

Zalesky, A., Fornito, A., and Bullmore, E. (2012). On the use of correlation as a measure of network connectivity. *NeuroImage*, 60(4):2096–2106.

Zemel, R. S. and Carreira-Perpiñán, M. Á. (2005). Proximity graphs for clustering and manifold learning. In *Neural Information Processing Systems (NeurIPS)*, pages 225–232.

Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., Stanley, H. E., and Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, 10(1):1–11.

Zügner, D., Akbarnejad, A., and Günnemann, S. (2018). Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2847–2856.