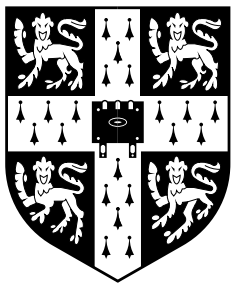
A Multilayer Approach to Geo-Social Networks: Models, Metrics & Applications

Desislava H. Hristova



Murray Edwards College
University of Cambridge

March 2017

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the regulation length of 60 000 words, including tables and footnotes.

A Multilayer Approach to Geo-Social Networks: Models, Metrics & Applications

Desislava H. Hristova

Summary

As technology becomes an inseparable part of society, the intersection between online human behaviour and offline physical presence fuels new aspirations of quantifying the human condition. The present dissertation explores the interconnectedness of these geographic and social levels of interaction through a multilayer network approach and demonstrates the knowledge that can be extracted about geo-social dynamics through such paradigm.

Leveraging rich data captured from mobile systems and online social networks, the advantage of using additional data layers when interpreting social and geographical systems is demonstrated. Bridging the theory and practice of multilayer networks, a minimal multilayer model is applied to six physical and digital international networks, including the international postal network which is analysed here for the first time. This model is further explored throughout this thesis in light of both online and offline social networks of varying scales, offering a simple but effective approach to multilayer data as such. In relation to urban dynamics, an interconnected multilayer geo-social network model is presented, capable of modelling entire cities as networks of people and places and offering powerful insights into urban development processes.

Along with each chapter and model in this dissertation, a set of application-specific metrics are discussed. While some of these, such as the global degree and overlap are extensions of single-layer network measures, others, such as the multiplex tie strength, translate between social theory and computational approaches for the first time. Some key theories explored are those of media multiplexity, homophily, social capital and brokerage, where the potential to under-appreciate these phenomena on the single-layer perspective is demonstrated. Furthermore, this new perspective on social and urban theory is shown to produce a number of valuable insights on how geo-social systems function in unison and specifically how social properties can be attributed to places and in turn how places shape social relationships. As an interdisciplinary body of work, this dissertation puts forward evidence in support of multidimensional data-driven studies of geographic and social systems as applied to fundamental questions in social science and urban studies, substantiating the merits of a multilayer approach to geo-social networks.

Acknowledgments

At the end of what has been a great chapter of my life, I am overflowing with eternal gratitude for those who have played a part in it. First and foremost, I am indebted to my supervisor, Cecilia Mascolo, who has been like an academic mother to me - never letting me fall off track but always giving me the freedom to pursue my curiosity. I will always carry with me the example and advice she has given me as I head for new horizons.

I am further grateful to Mirco for supporting me with his expertise and being an admirable academic example in those first two years when I needed it the most, and later to Matt who joined the effort. Thanks also to Jon Crowcroft and Pietro Liò, who gave me the confidence that I am putting together something that will result in a thesis along the way and Nishanth Sastry who helped confirm so following my viva voce examination. I have also been fortunate to have the mentorship of Pietro Panzarasa, David Lieben-Nowell, Daniele Quercia, Luca Maria Aiello, Alex Rutherford and the entire UN Global Pulse team. I am humbled by the great work they do and for having the chance to be a small part of it. The introduction into the wondrous world of research, however, I owe to Licia Capra way back at UCL, who also suggested I should do something called a PhD.

I would not have enjoyed this era so much if it were not for my friends, new and old alike. I am lucky to have had Alessandro, Lorena, Petko, Chloë, Neal, Christos, Sarfraz, Sandra, Deborah, Andrea, David and Xiao by my side during my time at the Computer Lab. I am grateful for the coffees, happy hours and dinners we have enjoyed together and the support we have given each other as friends. To my officemate and partner in crime, Tassos, I will be forever grateful for the moral support, location-based wisdom and comic relief. I feel so blessed for having been part of the Computer Lab community.

To my family and friends who have always been there for me, especially my two pillars of sanity - Sahand and Maria - I cannot imagine this journey without you. My parents - Maria and Hristo and my sister Evgenia, as well as Niki and Stefan are the reason I am here and I will never give up striving to make them proud. I am so fortunate to have had my stars Petya, Hanna, Roxana, Ina, Emma, Siyana, Malina, Vyara, Karishma, Alessandra as well as my baby stars Oona and Fanni and my entire BRC family shining over me for the past four years. I can't wait to share more adventures and endless conversations on how to change the world together.

Finally, I can't help but smile as I imagine how happy and proud my grandparents Sofka and Decho would be right now. They never travelled the world or went to school much but knew everything and so all of this is for them.

"Forty-two!" yelled Loonquawl.
"Is that all you've got to show for seven and a half million years' work?"
"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

- Douglas Adams,
The Hitchhikers Guide to the Galaxy

Contents

Chapter 1

Introduction

Almost all real and virtual systems are inherently composed of multiple layers or subsystems, which contribute to the wholeness of their functionality but can also be considered as systems in their own right. For instance, people in a city interact on a social level, but they are also embedded in geographical space, thus forming a *geo-social network* from these different layers of interactions. Network science has been largely successful in abstracting meaning from single-layer subsystems [11, 132], such as social interactions alone, and it is only recently that *multilayer networks* [96] have become a popular paradigm for the modelling of interrelated subsystems and entire systems, carrying the aspiration that these multilayer models can help us understand the bigger picture more realistically.

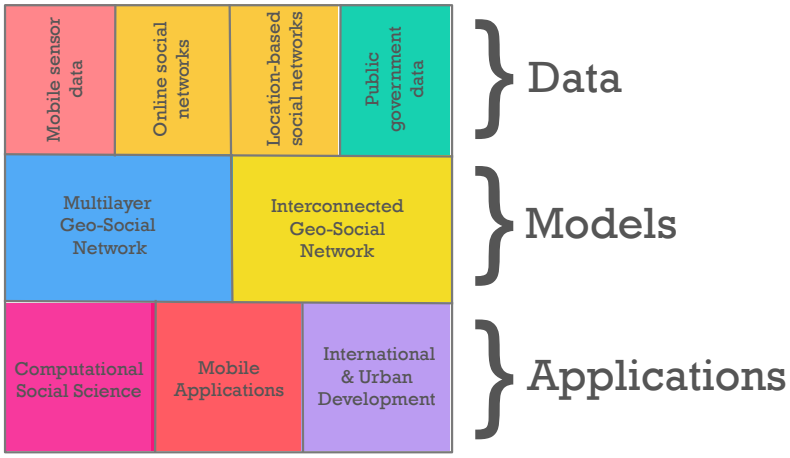


Figure 1.1: Dissertation structure: data, models and applications.

With the emergence of mobile systems capable of capturing the daily patterns of movements and the social life of people, Computer Science has been placed at the forefront of advancing our understanding of complex human behaviour. Unlike theoretical and survey-based approaches in other disciplines, empirical data-driven computational methods can be applied to quantify our theoretical understanding of complex systems.

An overview of this thesis is provided in Fig. 1.1 where the types of data, models and applications are illustrated. The introduction laid out in this chapter will set the stage for the rest of the dissertation and touch upon fundamental concepts which will later be explored in depth. The main focus here will be on the application aspects, while Chapter 2 will go into the details of multilayer models and methodology. Most importantly, this chapter introduces the technology and data which make this research possible. Starting from an overview of what systems and applications provide in terms of geo-social data and continuing to a summary of literature on human mobility and social interactions, this chapter concludes with a review of potential implications for computational social science, mobile applications and policy. At the very end, the thesis of this dissertation and its substantiation is specified through an outline of the structure and contributions of each chapter.

1.1 The State of Geo-social Network Analysis

The ubiquity of smartphones in urban areas, along with sensor-powered applications, have presented new opportunities to study the parallel social and geographical interactions of people at an unprecedented scale. In particular, location-based social network services have emerged as an invaluable source of such multidimensional data because of their wide adoption. These mobile applications entice people to share their location with friends and strangers, providing two of the essential ingredients in geo-social analysis: a social network of ties and a location map of their visits. On the other hand, rich interaction data from mobile sensing applications has opened up a new horizon for quantifying the human condition, from mapping how happy we are in different parts of the city [115], to monitoring global communication patterns across the world [80]. Moreover, due to a revolution in the quantification of services and open data, many other fundamental aspects of human global activity such as post, trade and transportation have consequently become available as data sources. These technological advancements form the backbone of the research presented in this dissertation, where a new paradigm for exploring such data will be introduced.

Social interactions and human mobility have been studied separately in a long tradition of scholarship, where fascinating insights have been drawn on the way we form relationships [79, 123], the way we interact within space and time [97], and the way that this shapes our opportunities in life [37, 63]. More recently, with the help of mobile data, striking relationships have been observed between these two domains, which render them inseparable and interconnected. Geography and mobility have been shown to play a fundamental role in the formation and maintenance of ties [139] and the social network of people can be inferred from their call data records [65]. Just recently scientists have also demonstrated that, inversely, mobility patterns can also be inferred from social interac-

tions where the same scaling laws have been shown to dictate both [58].

Although distance has frequently been viewed as an obstacle to social interaction, the "death of distance" [41] predicted through the advancements in communications and facilitation of long-distance travel has become an obsolete idea. As online social networking services mature, it becomes clear that distance is very much still alive and plays an important role in online tie formation [107, 127]. In the developed world, online social networking services have now reached a point of maturity where the majority of the online adult population (76%) was active on at least one of the most popular services in 2014 [61]. Nevertheless, online communication has been shown to play an important role in maintaining close and distant relationships, while decreasing the cognitive cost of doing so [62]. As part of the phenomenon of globalisation, it appears we have been growing increasingly connected through weak ties, from "six degrees of separation" in Stanley Milgram's 1967 small-world experiment and four degrees through Facebook topological shortest paths in 2012 [7].

Network science has pervaded the field of geo-social analysis due to its generalisable and intuitive models. Although social and geographic networks have mostly been considered in isolation, the recent popularity of multilayer networks [96] has presented an underexplored opportunity for modelling geo-social networks as systems composed of multiple network layers. This paradigm allows for maintaining the network properties of isolate layers independent while also as part of a system, where an interaction on one layer can have a cascading effect across layers. In the context of geo-social systems, this means that proximity-based interactions in physical space can be manifested as online social links and vice versa. Furthermore, the *multiplexity* of social interactions as a unified experience of online and offline encounters can be measured though such models, and their novel network configurations can reveal global connectivity and resilience, invisible from the single-layer perspective.

In fact, sociologists have long called for multidimensional studies of social phenomena such as homophily, or the tendency to form preferential ties with those similar to us [123] and the value of social networks [37]. Although still scarce, multilayer geo-social data has become increasingly available as more social networking services have become location-aware in the past few years, largely due to the business potential of location-based services. The systemic representation of social and geographical interactions as a multilayer network is extremely promising for the next-generation of recommendation systems and mobile applications, as well as fascinating from a sociological perspective. Modelling the dynamics of people-place interactions can also reveal novel insights about the social roles of semantic locations in the urban context and the wellbeing of neighbourhoods as will be further substantiated in this thesis.

1.2 Geo-social Networks in Practice

Geo-social networks as constructs of social and geographical interactions are in practice information layers made available by the data collection system. Historically, it has been impossible to obtain this level of granularity at a large scale and therefore most studies of geo-social human interactions have been restricted to small sample sizes and have been largely observational. The idea of studying the social relationships and movements of individuals in a sociometric manner had been introduced in 1934 by psychiatrist Jacob Moreno who explained the mobility dynamics of a group of runaway girls through studying their social network at school [128]. In 1948, the first larger scale study (1050 participants) of how geography, and in particular urbanisation, affects social tie strength was conducted in a survey-based fashion across Californian neighbourhoods by sociologist Claude Fischer [70]. Now, equipped with data only envisioned in the 20th century, there are new opportunities to study geo-social interaction dynamics in a systemic and multi-dimensional manner. There are broadly three types of geo-social data available at scale: mobile call and proximity sensor data, online social and location data, and quantified service data from technologies such as RFID.

1.2.1 From mobile systems to geo-social interactions

With the development of proximity-based sensors deployable on consumer mobile devices, such as Bluetooth technology, the dimensionality of human interaction studies has flourished. The appeal of using Bluetooth devices comes from its proximity-based wireless protocol, allowing for direct inter-device communication at a short range (usually up to 10 meters). Their scan discovery protocol identifies nearby devices using their unique identifiers and therefore serves as an excellent co-presence detector. Augmenting this proximity data with social network information is non-trivial, since device identification does not imply personal identification. Therefore, dedicated experimental projects need to be set-up for practical and ethical reasons. One of the most successful such projects is the MIT Reality Mining project [64], where personal and interpersonal behaviour was tracked in order to study the social dynamics of the student community. Other research has attempted to bring sensor-based geo-social analysis to the city level, where the "pervasive infrastructure" of cities was taken into account as a systemic view of human, space and technological factors and human encounters can be quantified in such setting [98]. At this level, however, there is no fine grained information available about individuals, while deployment still remains a challenge due to the large amount of observation and infrastructure requirements.

An undoubtedly large-scale complementary source of geo-social data that comes from mobile devices is Call Detail Records (CDR) from telecommunication providers. Using cell tower signals, the location of individuals can be approximated at the initiation of each

call or message. With the release of a few such datasets to researchers in the past decade, several notable studies have resulted in great advances in human mobility studies [78, 170, 139]. While certainly pervasive, large-scale and an important artefact of modern communication, CDRs are only approximate in terms of location and speculative in terms of social relationships. Furthermore, access to CDRs has been a privilege of few researchers due to the sensitive nature of the data and its proprietary nature.

The network assisted Global Positioning System (GPS) is currently the most accurate source of location data, on average up to a few meters, and along with location signals from WiFi and cell towers, provides reliable location positioning for mobile apps on smartphone devices. Perhaps the most viable source of geo-social interaction data, based on such location technology, are location-based social networks such as Foursquare, where users can "check-in" to semantic locations such as schools, cafes, restaurants and parks, broadcasting their location to friends on the application. In terms of data, this results in a proxy for human mobility augmented with a social network of friends. Many researchers have already taken advantage of such data, demonstrating the potential of augmenting social networks with geographic data for studying place-focused communities [31], friend recommendations [164], social forces such as homophily [201], and where people will check-in next [44]. Some models have tried to couple mobility with social contacts [180], with the goal of understanding mobility through social interactions and more recently it was demonstrated that the two are in fact driven by the same scaling laws [58], which invites for research beyond cross-sectional studies and towards novel applications of the interconnected nature of the two.

1.2.2 The online social network ecosystem

The online social media space, has been increasingly alluded to as an "ecosystem", due to its multitude of platforms, all competing for the same resource - user attention [6]. There are many other systemic properties in the ecology of online social networks, such as their intention to serve different purposes yet cohesively integrated in the World Wide Web. Some major segments of social media services include interest-based communities such as Pinterest, blogging platforms such as Twitter and Tumblr, photo-sharing such as Instagram and Flickr, location-based such as Yelp and Foursquare, and the all-prevalent Facebook. All of these services support social networking and are integrated to various degrees with each other, where content from one can be shared to another, contacts can be imported and accounts linked. This ecosystem has been understudied due to its recency, the lack of multi-platform data and models of analysis.

Nevertheless, some recent efforts have been made to understand the way users present themselves across platforms [91], the way they share content [141], and how traces from one service can be used to predict identity on another [77]. The access to multi-platform user activity and networks is very appealing from a commercial standpoint, as companies

compete to better match users to products, as well as from a privacy and security perspective. Efforts to study urban and social problems from a multi-platform perspective using the digital footprints of users, however, are rare due to their interdisciplinary nature and aforementioned constraints on this type of data. Furthermore, while mobile networks are pervasive, social media use is not and there are many considerations such as the validity and representativeness of such data [181]. These limitations are discussed in the context of each dataset used in this thesis.

While multi-platform studies are not yet widely popular, recent years have seen an elaborate body of research dealing with both the geographical and social aspects of online social networks. This is largely due to the shift towards geotagged content across social networking platforms and the popularisation of check-ins as a form of online content. On the one hand, location-based social features have been demonstrated to have vast potential for link prediction in online social networks [50, 164] and venue recommendations [136], while on the other hand, studies have shown the importance of distance in location-based social networks [163] and the gravitation laws that dictate universal properties in human mobility patterns [135]. Further to being excellent sources of geo-social data, location-based social networks have a strong link to urban computing, where they have been shown to be useful in neighbourhood modelling [49] and indicative of deprivation in neighbourhoods [152], among other applications. In this sense, location-based social networks are more than just products of available mobile technology. They are also self-contained ecosystems of physical presence and digital interaction. The potential of studying the urban landscape through modelling location-based social networks as systems of social and geographical interacting networks has been largely untapped.

1.2.3 National and international geo-social networks

While much of our insight into geo-social dynamic comes from the online world, there is a vast potential for alternative data sources as more and more global services become quantified through technologies such as RFID for tracking physical items in space. Major studies on urban underground transportation [74], global air transportation [81] and international trade [85] have already shown the potential of physical networks to model global dynamics in space. The notable work of Eagle [63] , demonstrated the potential of CDR interactions in the UK to explain deprivation through social and information metrics of diversity. In terms of multilayer studies, research on multimodal systems and cascading failures, which are not possible on the single-layer, has been conducted [33, 43, 74]. However, there is a gap in the literature in terms of socially augmented physical networks, where both physical and virtual network layers are taken into consideration, especially in the context of global connectivity. This is largely due to the differences in collection methods of social and physical data.

Nevertheless, it is possible to couple such disparate datasets through geography as some

global studies have shown [172], where social media data from Twitter and Yahoo! has been used to map the global alignment of interpersonal communication as posited by Samuel Huntington's theory of world cultural divisions [89]. With the presence physical layer data such as trade, international transportation and Internet topology, combined with global digital communication, there is vast potential to advance our understanding of global connectivity and international relations beyond the single layer as will be proposed in the present research.

1.3 Potential Implications

The link between social theory, physics and computational approaches in understanding social and urban systems is still nascent, whereas the potential for gaining novel insight from their combination is vast. The multilayer approach to geo-social networks put forth by this thesis makes use of several existing data sources in novel ways to revisit sociological and urban theory and apply it to new and old problems in Computer Science, Computational Social Science and Urban Studies.

1.3.1 Computational Social Science

The social theory of homophily examined by Miller McPherson et al in their seminal work *Birds of a Feather* [123] and the *Strength of Weak Ties* by Mark Granovetter [79] are perhaps the two most influential sociological works of the past few decades. Their impact has been profound on social computing where homophily has been extensively researched in online social networks in terms of tie formation [2, 3, 201] and weak ties have been shown to be more prevalent online, leading to a revolution in global connectedness and what came to be known as "networked individualism" [21]. Despite Dunbar's cognitive limit of 150 comfortably maintained social connections [176], online friendships have provided an easy and minimal effort way to connect with weak ties, and the more diverse they are the more social capital one can potentially generate [39]. Although these social forces are relatively well understood from a monodimensional perspective, the way in which people connect online and offline, both through social interactions and geographical ones is still not formalised extensively. One of the goals of this thesis is to bring dimensionality to social theories, which is something called for by McPherson himself [123] and show how this dimensionality adds to our understanding of these phenomena and their applications.

The idea of *multiplexity* or the multi-channel communication in social networks, however, is not new. Multiplexity in social networks has been observed and used to explain social dynamics from Renaissance Florence [142] to the Internet age [84] but not many generalisable models or applications exist at present. In her extensive research on the topic,

Caroline Haythornthwaite, conducted a study where she asked a group of university members to disclose the type of relationships within their social network, their frequency of communication and media used [83]. She coined the term *media multiplexity*, where she posited that media usage varies between weak and strong ties and that strong ties tend to utilise more of the available communication channels, showing that the medium is indeed the message. The present dissertation operationalises and applies this principle to online geo-social networks and problems of link prediction in the following chapters.

1.3.2 Mobile Applications

Location-based services have added a significant advantage to mobile applications by bridging the online and offline realities of users. While some location-based apps such as Google Maps and Yelp focus on local search, others have been built on social networking capabilities such as Foursquare. Furthermore, Facebook has location-based capabilities through Facebook Places and others such as Twitter and Flickr allow users to geotag their content. From a service provider perspective, this allows for context-aware recommendations and more detailed user profiling as well as being better equipped to face the challenges posed by content-delivery networks [162].

Although location-based social networks have resulted in an invaluable source of geo-social interaction data, they struggle to remain competitive in the online social network ecosystem. Due to the lack of clear business models and failure to engage users in a sustainable manner, location-based social networks such as Brightkite and Gowalla have become obsolete while market leader Foursquare has struggled for years to establish a niche in the location app space. Research which has tackled the death and "autopsy" of online social networks [75], has identified the weakness of social resilience in social networks as the reason for their demise. It is therefore crucial to understand the driving forces of tie strength and the role of location in such services. Mobile systems can further benefit from a multilayer approach with regards to application design, where a systemic perspective of the geographical and social components of the system can be seen as interacting yet distinct could help establish unique competitive advantages and long-term sustainability.

Recommender systems such as those based on link prediction algorithms for online social networks and venue recommendations are an integral part of online services. The present thesis will add to the analysis of location-based services by examining them from a multilayer perspective and will explore new potential applications to "socially-aware" place recommendations where the social roles of places are taken into account. Furthermore, applications to urban development such as tracking human mobility for the identification urban migration patterns will be discussed in light of the future generation of "socially-aware" mobile systems.

1.3.3 Urban Development

The city has been rightfully been alluded to as an organism, due to its complex dynamics and processes [16]. Much imagination and research has already been poured into the vision of a "smart city" governed by computer-supported systems. A part of this new science of cities is the quantification of urban services such as public transportation [168] and human mobility modelling [78]. Augmenting this understanding with a systemic approach to urban modelling can help urban planners and policymakers make decisions based on knowledge of the whole ecosystem of services.

Another area of interest is the measurement of critical indicators and understanding inequality on the neighbourhood level. At present, government census studies are the most widely used measures of urban socioeconomic wellbeing at the urban level. However, recent works proposing the use of user-generated content and social media for measurement have emerged [111, 188, 151, 184] based on the availability of these new forms of data. Using interim real-time measures of the urban pulse is appealing from a temporal and cost perspective but can be challenging due to the demographic biases of digital media users [61]. Nevertheless, as this dissertation will show, there are also opportunities in tracking the social media population, which can lead to change in urban policy regarding urban renewal and gentrification.

1.4 Thesis and Substantiation

The goal of this body of work is *to analyse and interpret multilayer network models of geo-social systems by applying them to real world problems through a data-driven computational approach.* As laid out in the preceding sections, social and urban theory has been explored mainly in a unidimensional manner and many sources of rich multilayer data have been untapped to such ends. Therefore, the thesis of this dissertation is that *a multilayer network approach to urban and social theory can advance our understanding of geo-social dynamics beyond what is possible from studying their social and geographical components in isolation.*

The substantiation of this thesis will be incremental - starting with the exploration of a minimal multiplexity model for measuring global connectivity and socioeconomic similarity using a number of physical and digital networks. This will be followed by an expansion to small scale offline and large scale online networks with the goal of understanding tie strength, media multiplexity and homophily in the multidimensional setting. As a demonstration of the interconnected nature of people and places, we will also discuss a multilayer interconnected network model where novel insights will be shown in the domain of urban computing and social capital. In the context of each research application, a different set of tools and techniques will be utilised to present multiple threads of analysis evaluat-

ing the benefit of multilayer network modelling and analysis in comparison with current research and "traditional" single-layer models. Each chapter will present a number of structural and interaction metrics and add to the original multiplex model in light of different applications as outlined next.

1.5 Contributions and Chapter Outline

This dissertation makes several novel contributions: firstly, to the field of complex networks where multilayer models have recently enjoyed a renaissance, however, most work done in the field is theoretical at large, while this dissertation provides empirical and data-driven modelling techniques and evidence; second, to the field of computational social science, where multilayer network models and metrics from a geo-social perspective have not been applied; third, to mobile systems and computing, presenting novel ways to conceptualise location-augmented systems where data has largely been utilised in a unidimensional manner; and finally to the field of urban computing, where interconnected networks of people and places have not yet been introduced despite the great potential to revolutionise our understanding of space and its social uses. The state of the art on multilayer models and current applications will be outlined in Chapter 2, while the contributions made in this dissertation are detailed in the following chapters:

- **Chapter 3: Global Multiplexity**: In this chapter, a minimal *multiplex model* is presented as a collection of graphs. The idea of multiplexity, or multi-level communication, is explored in the context of this model, using six distinct sources of global interaction data: the global trade network, consisting of the value of exports and imports made between countries; the global migration network where the estimated migration flows of people around the world are reported; the global flight network where the number of flights between countries is recorded; the global IP traceroute network, constructed on top of the Internet topology; the digital communication network derived by Twitter mentions and Yahoo! emails on a global scale; and finally the international postal network, which is analysed in this dissertation for the first time. These physical and digital layers of online and offline international resource flows are combined and studied together through the multiplex framework, where it is shown that the global degree is useful in approximating critical socioeconomic indicators. Further, the *community multiplexity index*, a measure of community membership, similarity is applied to approximating the socioeconomic profiles of countries, where it can be used to estimate critical values for international development purposes when such data is missing for a particular country.

- **Chapter 4: Multiplexity and Social Ties**: This chapter further substantiates the role of multiplexity in geo-social networks by formalising a *multiplexity weight*

following empirical analysis of a student community geo-social dataset. The social relationships of students and their level of communication multiplexity is analysed in terms of proximity, phone calls and text messages, where more communication channels were shown to be strongly indicative of a close friendship. This concept is further applied to studying homophily in the community where students with multidimensional ties were a lot more likely to be similar in terms of music taste, political orientation and various other dimensions. We further explore similarity in terms of structure and interactions on a large-scale using Twitter (social) and Foursquare (location) data. A number of features for link prediction are presented in this multilayer context, some of which extend existing single-layer metrics such as the *multilayer Adamic/Adar* coefficient, while others use interactions from the two heterogenous network layers such as the *global similarity*. Using three testing and training sets of location-based, social and mixed features, we observe the superior performance of multilayer mixed features to predict multiplexity in online social networks using a supervised learning framework.

- **Chapter 5: Social Diversity in Geo-Social Networks**: In the final research chapter of this dissertation, the concept of social capital is explored in terms of structural diversity in the multilayer network. First, a quantitative measure of social capital, or brokerage, is applied to the multilayer setting in a two-layer online social multiplex consisting for Twitter and Foursquare layers. It becomes empirically evident that by considering just one layer - social or geographic alone - social capital can be under or over-estimated. The concept of brokerage is then redefined for places, where an interconnected network model of people and places is introduced and the *brokerage power of a place* is defined in terms of redundancy in the social network of its visitors. Three other metrics of the social diversity of places: entropy, homogeneity and serendipity are introduced here. Entropy is the predictability of visits made by visitors to a place, homogeneity is the visitor diversity in terms of characteristics and serendipity measures the probability of the social composition of a place in terms of visitors. These measures of diversity are then compared to indices of deprivation for London, where a positive relationship is identified. Neighbourhoods which are characterised by high social diversity but have high deprivation, are identified as neighbourhoods which are undergoing processes of change such as gentrification.

The final chapter in this dissertation, reflects on the findings and applications of this work and provides a vision for future research.

1.6 List of Publications

The following academic publications make up the research conducted towards completing the present thesis. These consist of my sole efforts and technical contributions but would not have been possible without the support and guidance provided by my co-authors.

Chapter 3: Simple Multiplex Model for Complex Network Analysis:

- Hristova D, Rutherford A, Anson J, Luengo-Oroz M, Mascolo C. The International Postal Network and Other Global Flows as Proxies for National Wellbeing. PLOS ONE, 11(6), 2016.

Chapter 4: Multiplexity and Social Ties:

- Hristova D, Musolesi M, Mascolo C. Keep Your Friends Close and Your Facebook Friends Closer: A Multiplex Network Approach to the Analysis of Offline and Online Social Ties. In Proceedings of the 8th International AAAI Conference on Web and Social Media (ICWSM'14), 2014.
- Hristova D, Noulas A, Brown C, Musolesi M, Mascolo C. A Multilayer Approach to Multiplexity and Link Prediction in Online Geo-Social Networks. EPJ Data Science, 5(1), 1-17, 2016.

Chapter 5: Measuring Social Diversity in Geo-Social Networks:

- Hristova D, Panzarasa P, Mascolo C. Multilayer Brokerage in Geo-Social Networks. In Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM'15), 2015.
- Hristova D, Williams M, Musolesi M, Panzarasa P, and Mascolo C. Measuring urban social diversity using interconnected geo-social networks. In Proceedings of the 25th International Conference on World Wide Web (WWW 16), pages 21–30, 2016.

Other Publications:

- Hristova, D Liben-Nowell, D Noulas, A Mascolo, C. If You've Got the Money, I've Got the Time: Spatio-Temporal Footprints of Spending at Sports Events on Foursquare. CityLab Workshop. In Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM'16), 2016.

- Hristova D, Quattrone G, Mashhadi A, Capra L. The Life of the Party: Impact of Social Mapping in OpenStreetMap. In Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM'13), 2013.

- Hristova D, Mashhadi A, Quattrone G, Capra L. Mapping community engagement with urban crowd-sourcing. When the City Meets the Citizen Workshop. In Proceedings of the 6th International AAAI Conference on Web and Social Media (ICWSM'12), 2012.

Chapter 2

Multilayer Networks

The following chapter gives a general introduction to multilayer networks and metrics with comparison to their single-layer counterparts. Network representations have emerged as an extremely powerful and general framework for analysing and modelling systems as diverse as transportation, biological processes, academic authorship and logistics among others [11]. Network science provides powerful tools for understanding such systems with large sets of coupled components and emergent behaviours more generally known as complex systems. Despite the broad range of applications, traditionally, most network problems have been examined with a single graph representation, largely ignoring interconnectedness and parallel interacting networks.

From multimodal transportation systems to interpersonal relationships, most real world entities are connected in more than one way. One example is social network analysis, where human relationships have been extensively studied as a formation of a single type of personal interaction, oversimplified usually as friendship bearing a weight proportional to interaction. However, people interact for various overlapping reasons such as work, social, family and love life. This multiplexity, or the complexity caused by the existence of more than one link between entities, is ubiquitous across science domains concerned with systems.

In the past few years there has been a rekindled interest in unifying and formalising the multilayer paradigm, particularly in the domain of physics and biology due to the increased availability of suitable data for these models and their potential to represent systems more realistically. Networks with multiple edges of different types between nodes are most frequently referred to as network of networks [51], multiplex networks [189], multirelational networks [195], and, more generally, as multilayer networks. Although these models are all based on the same premise, that there are multiple relationships between entities in a system, terminology largely differs based on domain. Some existing examples include interacting systems such as the power grid [33], biological networks [22] and transportation in the urban context [74, 56]. In this section, we will review three of

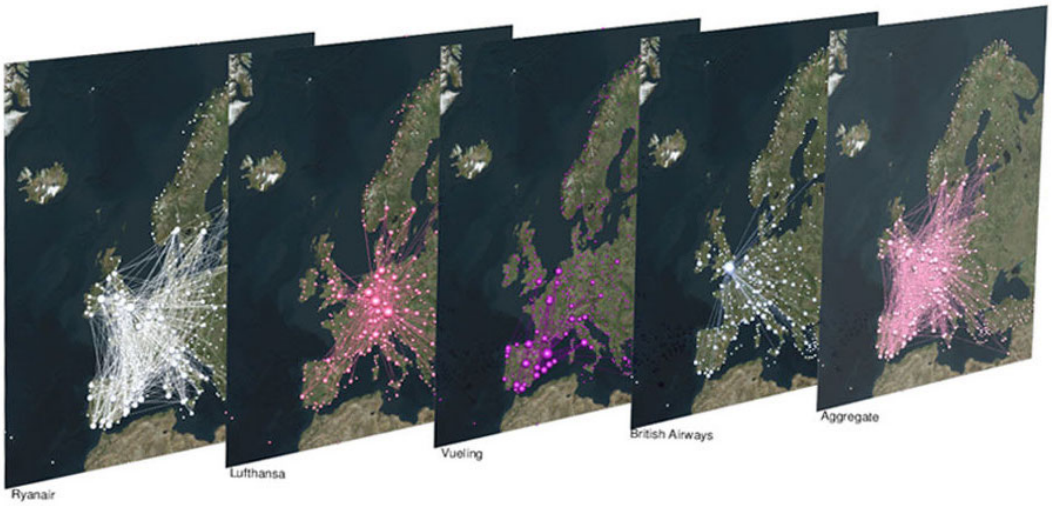


Figure 2.1: The European flights network [56].

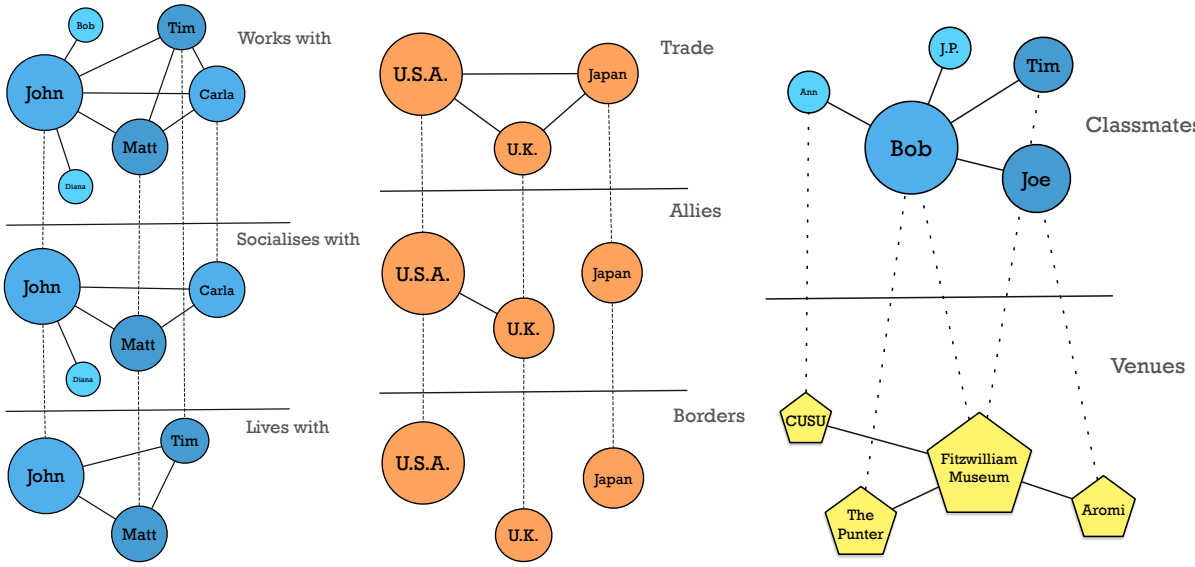
the most common multilayer models, which relate to the present dissertation along with their metrics and applications.

2.1 Multirelational Network Models

Networks are typically modelled as graphs composed of a set of nodes (vertices) and edges (links between them), along with certain fundamental properties such as directionality and weights. When multiple sets of edges exist between nodes, such as different modes of transportation connecting two stations, extensions to the network paradigm are necessary in order to capture this multirelational nature, although the fundamentals of graphs remain the same.

Multilayer networks can be represented as multigraphs, interdependent networks or rank-3 adjacency tensors, depending on their nature and the background of the researcher but they all have the common elements of nodes, edges and layers. The model which is adopted is largely dependent on the field and context of study. Recently, most of the multilayer literature has been produced from the physics domain, where many generative models have been proposed. However, with the growing availability of empirical data, it is increasingly important to also gain understanding of multilayer networks from a data-driven perspective.

One example of such data can be found in Figure 2.1, where the European flights network is shown as a system of different airlines and as an aggregate network, demonstrating the flexibility of multilayer networks to be studied layer by layer as well as together in a systemic manner. As will later be reviewed in this chapter, not all layers need be of the same type as in this example and there are many cases in which interactions happen in



(a) Multilayer social network

(b) Multiplex network of international relations

(c) Interconnected people-place network

Figure 2.2: Multilayer, multiplex and interconnected network examples.

various physical and digital dimensions and even between different types of entities. This section covers only general multilayer, multiplex and interconnected network models of multirelational networks relevant to this dissertation, however, more complete surveys of generative and theoretical models can be found in [96] and [22].

2.1.1 Multilayer Network Models

The notion of a *multilayer network* is the most general when it comes to studying multirelational networks. The only implied requirement is that there is more than one network layer of nodes and edges. This generality has led to many differing representations. In this section we will discuss the two most popular: graph based and tensor based.

The adjacency-tensor paradigm stems from Mathematics and is a vector-based representation of multidimensionality in high orders. For example, in [55] the authors extend the single-layer tensor representation to a multilayer one, where multiple types of relationships can be represented. This approach requires the definition of two second order adjacency tensors - one which represents relationships within layers and one between them. The multilayer representation is then derived from the product of these two tensors. For example, $\mathcal{A}_{uv\alpha\beta} = \mathcal{A}_{uv\alpha_1\beta_1\cdots\alpha_d\beta_d}$ represents the order-2 adjacency tensor where element $\mathcal{A}_{uv\alpha\beta}$ has a value of 1 if and only if nodes u, v in layers α, β have an interlayer link. This adjacency tensor is the product of the layers and vertices in the multilayer set-up.

On the other hand, the same properties can be obtained in a graph representation: $M = (V, E_M, L)$ where the standard graph notation $G = (V, E)$ is extended to include the set of layers L. If nodes are different across layers, an additional set V_M can be included

which is more convenient than in tensor representations where missing nodes need to be 'padded', although this can be overcome with tensor flattening [55]. This model is general enough to represent networks with varying topologies, time and other dimensions. With certain constraints, it can be used to represent networks that are multiplex, edge-coloured and sequential, as these models are all special cases of networks with multiple layers. A very simple multilayer network as an example is the social network in Fig. 2.2a, where different types of interpersonal relationships are ascribed to individuals. Some people do not appear in all layers because they do not interact in that particular layer in the multilayer network.

2.1.2 Multiplex Network Models

Multilayer and *multiplex networks* are often used interchangeably because a multiplex model is the most common manifestation of a multilayer network. It is more specialised in two respects: a) the nature of its interlayer links and b) the replicability of its nodes across layers. In [55], the authors specify that "A multiplex network is a special type of multilayer network in which the only possible types of interlayer connections are ones in which a given node is connected to its counterpart nodes in the other layers". In many cases, this means that nodes should be present in all layers, and therefore the network is *node-aligned* but this is usually not a requirement.

Multiplex networks are perhaps the only multilayer construct that can be easily translated into a monoplex single-layer network in a straightforward manner. This of course is possible only for non-ordered and non-weighted interlayer edges. This aggregation is useful for applying standard network metrics to multiplex networks. Usually the weights of the same edges across layers are linearly combined, which yields a weighted graph [19]. In addition, each layer can have a different weight encoded as a coefficient. In some cases, however, this information is purposely disregarded in order to examine the connectivity in the multiplex in itself [15], although in many scenarios it has been argued that a full multiplex model has a theoretical information advantage over aggregated and single-layer models [124, 54, 57, 134]. An example of a multiplex network is given in Figure 2.2b, where all nodes are present across layers and each interlayer edge represents the coupling between the same node across layers given three types of international relations between countries. In the layer "borders" an edge case is given where neither of the nodes are connected with each other but they are nevertheless embedded in this geographical layer.

2.1.3 Interconnected Network Models

Interconnected networks have many variants across different fields of study, including urbanism [51], engineering [110] and epidemology [161], and are some of the most practical and widely applied models of multilayer networks. This is due to the fact that, in many

cases, they model interdependence in systems and this can lead to novel understanding of resilience and cascading effects. In [33], the authors used data from the 2003 Italian electrical blackout, modelling the series of cascading failures which occurred in relation to power stations presenting an interconnected network of Internet servers and the geographical locations of power stations. This was followed by work done on power grid cascading failures to predict the cascades and aim to minimise risk from such failures [32]. As in this example, in many cases interconnected networks have different sets of nodes on different layers, similar to node-coloured networks, also referred to as multitype networks [187].

In terms of representation, the only constraint that interconnected networks have is that they are *layer disjoint*, or that each node exists only in one layer [96]. In this sense, they are very similar to node-coloured graphs, where each node has a colour analogous to a layer in the multilayer paradigm. Because interconnected systems are often interdependent, each layer can be represented as a graph in and of itself, along with a mapping of their interdependency. For example, in [33], where on one layer there is the geographical network of power stations in Italy and on the other the network of connected servers to those power stations, the relationship between the two networks was defined as a bidirectional link $A_i \longleftrightarrow B_i$ modelling the fact that if a node in one layer fails all nodes for which the bidirectional link is true will fail too, leading to a cascading failure. Although less critical, a visual example of such networks is given in Fig. 2.2c, where a social network of classmates and the place network of the venues they have visited are interconnected and the relationship between the two can be studied.

2.2 The "New Physics" of Multilayer Networks

Most multilayer metrics and models aim to extend traditional network science concepts to cater for the presence of more than one network layers. We will discuss several of the most important concepts of these "new physics" of multilayer models related to this dissertation – starting from their traditional monoplex definition and extending into their multiplex manifestations.

2.2.1 Degree

Perhaps one of the most fundamental metrics in network science is the node degree, which measures the importance of a node in terms of number of links in the network. This is the number of adjacent edges and consequently corresponding nodes that make up the focal node's neighbourhood. In traditional single-layer graphs, this concept has been further extended to consider directed networks by measuring the in and out degree of a node's incoming and outgoing edges and to weighted networks, where the edge interaction weights are added and sometimes normalised to produce the weighted network degree.

A crucial element of defining the degree of a node is the definition of the node's neighbourhood. In multiplex networks, where it is possible to aggregate the network into a single-layer graph, the traditional concepts of neighbourhood and degree can be easily applied. The most straightforward approach to measuring the degree of a node in the multilayer network is by considering the union or intersection of edges across layers. The aggregation, however, can be done in multiple ways and it is contextual to the problem at hand. For example, in [28] the authors define the multi-layer neighbourhood of a node to be the collection of neighbours who are adjacent on at least $\alpha(1 \leq \alpha \leq |L|)$ in the multiplex where α is the layer threshold value.

On the other hand, a vector version of a node's degree has been proposed [15], where the vector $k_i = (k_i^1 ... k_i^M)$ represents the multilayer degree of a node i across layers. Since this does not give a clear signal of a node's importance in the network, the overlapping degree needs to be computed:

$$o_i = \sum_{\alpha=1}^{M} k_i^{[\alpha]} \tag{2.1}$$

which is essentially an aggregation of the vector. The approach to multilayer degrees in this dissertation is similar to both of these methods, where a threshold for minimum layer connectivity is set and an aggregation is the performed in terms of weighting the edges. This is discussed at more length in Chapter 3.

Another approach has been to define a set of layers such that $D \subseteq L$ for which to consider all edges (u, v, d), as defined by the authors in [18]. This flexible definition allows for the exploration of various layer configurations and their relative importance, however, the number of possible sets of D grows exponentially with the number of layers causing difficulties in interpretation and computation. Another variant of this is the concept of *multilinks* introduced in [19], where a binary vector is used to store the adjacencies of pairs. The degree of a node is then the number of edges contained in the vector across layers and the weighted version is the sum of weighted edges on each layer. Apart from extensions of single-layer metrics, there has been a novel notion of a *multiplexity degree* which is concerned with the number of layers in which a node appears [160]. This however is only applicable to layer-disjoint models, where a node need not be present in every layer.

2.2.2 Triples and Triads

Beyond the concept of a node's neighbourhood, there is the connectivity between members of that neighbourhood, which includes edges between neighbours. In this expansion of the neighbourhood, special kinds of network motifs, called triangles emerge, where triads of connected nodes are formed. The number of such connected triads determine important properties of the network such as its clustering coefficient. The clustering coefficient is a

measure of the number of connected triangles over the total number of possible triangles, dependent upon the network degree in the global version and the focal node or ego's degree in its local network counterpart, determining how tightly knit an (ego)network is.

The definition of a triangle and consequently the network clustering coefficient becomes more complex as layers are added, since transitivity can extend to neighbours across layers as discussed in the next section on network distance. Several attempts to define clustering coefficients for multilayer networks have been made [15, 22, 27] but they are all dependent on the definition of transitivity and paths in the model. For this same reason, it has also been a challenge to bring a stable definition of a community and methods of community detection into the multilayer context. This dissertation makes use of traditional community detection techniques such as the Louvain modularity optimisation method [20] for detecting communities on the single layer. At the heart of this community detection method is the modularity function, which determines the strength of division between different clusters of nodes in the network. By optimising this function, groups of nodes are identified which are interconnected amongst themselves to a greater extent than to the rest of the network. Despite efforts to explore information similarity in terms of community structure between layers [90, 13] as a method for dimensionality reduction, the community detection literature with regards to multilayer communities is still nascent. Nevertheless, a few notable works have brought concepts such as modularity into the multiplex context [129] and a few other novel ways of community detection in multilayer networks [52, 147].

On the other hand, network motifs, where these triangles are incomplete are called structural holes. This concept is of great significance in social networks, where it relates to the brokerage power of a node and sequentially a structural competitive advantage in the network and a person's social capital. A measure of brokerage which is leveraged upon in Chapter 5 is defined in Burt's seminal work on structural holes [36]. Burt defines the effective size of a node's neighbourhood as the non-redundant portion of it. The effective size S_i of node i can be expressed as:

$$S_i = \sum_j \Big[1 - \sum_q p_{iq} m_{jq}\Big], \quad q \neq i, j, \tag{2.2}$$

where

$$p_{iq} = \frac{z_{iq}}{\sum_j z_{ij}}, \quad i \neq j \tag{2.3}$$

is the normalised weight of the link between nodes i and q over i's local neighbourhood where Z is the adjacency matrix and therefore if $z_{iq} = 1$ there is no structural hole and if $z_{iq} = 0$ there is a structural hole and no redundancy in the network in the case of an undirected weighted network, and where

$$m_{jq} = \frac{z_{jq}}{\max_k(z_{jk})}, \quad j \neq k \tag{2.4}$$

is the marginal strength of node j's link to node q (i.e., the weight z_{jq} of the link connecting nodes j and q divided by the maximum link weight node j has with any of its contacts). The value of m_{jq} can be reduced to z_{jq} in the case of unweighted and undirected networks in which the only weight a link can have is one. According to Equation 2.2, we have: $1 \leq S_i \leq k_i \ \forall i$, except for the case in which node i is an isolate (where S_i is set equal to zero). Thus, the brokerage of node i's neighbourhood ranges from the minimum value of one, when all pairs of node i's contacts are connected with each other, to the maximum value equal to the node's degree k_i (i.e., the number of links incident upon node i) when there is no link connecting any pair of i's contacts. What this essentially measures is therefore the number of non-redundant connections a node has to otherwise disconnected others. A simplified "unpacked" version of this was introduced by Borgatti in [23] as:

$$S_i = n - \frac{2t}{n} \tag{2.5}$$

where n is the number of neighbours of i and t is the number of links between them, excluding links to i. Structural holes have not been defined in the multilayer network setting yet and in Chapter 5 we make the first attempt to bring their classical definition into the multilayer context.

2.2.3 Network distance

Network distance is typically measured as the length of a path from one node to another. This may include weight and directionality trade-offs and many algorithms have been proposed to optimise for certain requirements such as length and cost. Depending on the real world context of a model one can define different notions of transitivity in multilayer networks. One such way is to define paths in terms of inter-layer connections, where nodes can form triangles across layers. In [46] the authors define this in terms of inter-layer steps, where at each step a walk can continue on the same layer using what is termed as *supra-walks*. This approach can be computationally intense for many layers, although it makes an important distinction between the triadic properties of social networks, which tend to be consistent across layers and the properties of transportation networks, which tend to form more inter-layer triads due to their multimodality.

The definition of distance in terms of walks and paths is an important concept on which many other network measures and algorithms rely, most importantly the shortest path definition. One way in which paths have been considered in multilayer networks, given the possibility to traverse between layers, is as the ratio of shortest paths between a pair of nodes across more than one layer over the total number of shortest paths, termed

interdependence of the multiplex [133]. This can be useful in understanding failures in transportation systems for example, where the cost of switching layers is also higher and should be considered in the path calculation if such cost exists. In this dissertation we use the notion of multiplexity to define the network distance, as will be further detailed in Chapter 4.

2.2.4 Layer overlap and correlation

One of the fascinating aspects of multilayer networks is the way in which their layers are connected with each other and as part of a system. Two ways to measure this relationship are the overlap between layers and their correlation in terms of degree, edge weights or other structural properties. The overlap is the similarity in terms of the number of common nodes or edges that two layers share. Bianconi defines the *global overlap* in the multiplex as the total number of edges two layers share [19]. Another way to compare edges is by looking at the correlation between weighted edges on different layers, or the adjacency matrices or two layers. Such analysis is extensively conducted on the International Trade Network in [12], where commodity-specific layers were correlated and hierarchically organised to produce a taxonomy of commodities in the multilayer network.

A significant portion of early multilayer network exploration has been dedicated to degree-degree correlations, also know as mixing patterns [131] or network assortativity. Typically measured as the degree correlation between two connected nodes, it indicates to what extent high degree nodes cluster with other high degree nodes. In the multilayer setting this mixing has mainly been measured in multiplex networks, where the degree of a node is correlated with itself on another layer [126]. Some results in the literature suggest that such strong correlations are in fact harmful for collaboration in certain types of social networks [194]. In Chapter 3, we will observe how countries carry their degree properties across networks of international relations and what this means in terms of edge correlations across networks as well.

2.3 Real World Applications

The majority of multilayer scholarship in recent years has been purely theoretical, exploring generative multilayer models through simulated experiments [96]. Although, this has been highly enlightening in view of the statistical mechanics of multilayer networks, there is a significant gap between theory and practice, where real world multilayer networks are not yet fully understood. Nevertheless, there are some notable examples of applications such as [60], where the authors use data from the National Centre for Environmental Prediction in Germany is used to analyse the atmosphere's vertical dynamical structure through interacting climate subnetworks. The authors in [34] used MRI and other brain

data to explore the relationship between structural and functional brain networks, leading to the discovery of significant interdependence between the two, and more importantly the brain's functions were found to follow the small-world topology properties much akin to spatial complex networks. In what follows, we will explore two domains of empirical multilayer studies which are most closely related to the present dissertation – that of online and offline social networks and that of geographical urban and global networks.

2.3.1 Social Networks

Social networks as well as other systems have been found to consistently exhibit a power-law degree distribution, small-world phenomenon, centrality and modularity [132]. Despite the success of social network analysis, the multidimensional nature of human relationships has been largely ignored in the past. Nevertheless, *multiplexity*, or the cardinality and type of edges, has been well accounted for in sociology. Sociological studies refer to multilayer social networks as "multiplex networks" since the 1970s [189], where kin, neighbour and coworker relationships are explored. Observational studies since the late 80s have reported that people who are linked through more than one way, have a stronger bond because they have more ways and reasons to communicate with each other [70]. In [159], Sampson (1968) describes several social relations among a group of men in preparation of joining a monastic order in an ethnographic study of community structure over time. Another example of small observational studies of human relations comes from [94], where the work and friendship interactions at a tailor shop in Zambia are analysed in relation to worker strikes. Other studies of social multilayer systems have been done also with small troops of baboons (n=12), where multiplexity is based on behavioural dimensions and perturbations in one dimension (layer) are measured across the social system [14]. With the popularisation of the Internet, the same was found to hold true of online media such as email, chat, and social network sites in [83], where the author studied the implications of multiple media usage on social ties in an academic organisation and discovered that multiplex ties (those which use multiple media) indicate a stronger bond.

It is easy to see how the multilayer paradigm fits social network analysis, where different relationships that are not mutually exclusive can be ascribed to the same two people (e.g., colleagues, friends, and siblings). The same can be stated of two people's communication (e.g., phone, email, and face-to-face); their interactions offline (e.g., meet, travel, and participate in sports) or online (e.g., tag a photo, re-tweet, and video chat). Despite the observable multilayer nature of online social networks (OSNs) as a system [96, 95, 26], there is little empirical work exploiting data-driven applications in the domain of multilayer OSNs, especially with respect to how location-based and social interactions are coupled in the online social space [141, 103]. One large-scale example applied to virtual social networks was presented by Szell et al. [177], where multirelational interactions rep-

resenting different relationships between the players in a massive multiplayer online game were used to model social ties and relate them to social balance theory (i.e., 'the enemy of my enemy is my friend'). Some other examples consider online social networks like Twitter [53], where the spread of information around the announcement of the discovery of the Higgs boson-like particle at CERN is explored through the multiplex interactions of tweeting, re-tweeting and replying to existing tweets; in [119] the authors studied different social networking services through the Friendfeed social network aggregator; and in [105] the authors collected multidimensional social network data based on Facebook to study the formation of groups in a university setting. This dissertation aims to extend the existing literature by modelling and analysing online and offline social networks through the multilayer models described in this chapter in a variety of contexts.

2.3.2 Urban and Global Dynamics

In transportation networks, spatial interacting networks have been used to analyse the structure of airport and railway networks in India, showing the effect of space in determining link probability [82]. The application of multilayer networks to the public and international transportation domains has been demonstrated throughly, as these systems are natural examples of multiplexes with multiple lines and modes along with a high level of interdependence and layer transfer cost [74]. Similarly, the London underground has been explored as a multiplex of transportation layers [56], where the authors address questions related to the efficiency and resilience of the system by exploring its interconnected nature through random walks.

International dynamics are also naturally multirelational. Work taking this into account has been done extensively in the international commodity trade domain [12] and global air transportation [43]. Both of these works demonstrate novel ways of exploring existing data, in the former case from the United Nations Commodity Trade Database, collecting trade data since 1990, and, in the latter case, by collecting data on European airlines and airports where rescheduling and resilience in the network were under investigation. Other examples include the global cargo ship network [93] as a separate system, as well as the worldwide coupled air and sea ports multilayer network, where it was found that the more inter-similarity there is the more robust the segment of the network is. A similar multimodal paradigm is explored in [82], where the dynamics of overlapping spatial networks are studied. One of the novel contributions of this thesis is the study of heterogeneous layers of global interactions, in particular physical and digital in parallel. This is further discussed in the following section where the contributions to literature and future outlook is presented.

2.4 Present Dissertation and Future Outlook

This chapter has reviewed the major trends in multilayer models, metrics and applications. With a plethora of terms describing this paradigm, we have seen some differentiations between three of the most popular mutirelational models - the generic multilayer model, the node-aligned multiplex model and the interconnected network model. The tendency for defining metrics for multirelational models has been to predominantly translate pre-existing single-layer metrics into the multilayer context of the model under hand, with few novel metrics specific to these models. Furthermore, the great majority of work in the field has been theoretical, with relatively few empirical examples.

This dissertation takes steps in exploring the introduced models from a data-driven computational perspective, where social and urban theory is applied to the systemic representation of multiple layers of data. Furthermore, along with validating concepts such as degree, distance and other traditionally monoplex metrics, new measures for exploring interconnected people-place networks will be introduced such as serendipity and structural holes. Most importantly, this thesis contributes to the scarce empirical literature on multilayer networks in geo-social systems.

Data-driven approaches to multilayer networks have mainly been applied in sociology, where small observational studies have attempted to model social dynamics from a mulitrelational perspective. As we have seen in this chapter, there is a general lack of studies which model data in a multilayer fashion, rather than try to fit data to formal models. In Chapter 3, we will demonstrate how six types of physical and digital international relationship types can be modelled collectively as a global multiplex network. From this network representation, we will show how the global degree and community multiplexity index can be derived to study the socioeconomic profiles of countries. Although some of the works reviewed in this chapter have touched upon metrics of multiplexity itself (such as interconnectivity), we show how in practice such fundamental metrics can be of great benefit to international development.

Furthermore, in Chapter 4, we will see how survey and mobile sensor data collected for the purpose of single layer analysis can be explored in novel ways to derive insight about social relationships through mutliplexity. This chapter also examines media multiplexity and homophily social theories in the new light of multiplex interactions. In addition, the applicability of multiplex models and metrics, including contexutally adapted gravity and multilayer interaction measures, is demonstrated in the domain of link prediction in online social networks. This work again demonstrates novel ways of adapting available data - from Twitter and location-based social network Foursquare, into an integrated geo-social multilayer network.

Finally, in Chapter 5, we will study multilayer brokerage for the first time as applied to two geo-social multilayer models. In the first, we study the online/offline paradigm

of generating social capital through brokerage, where the presence of structural holes is explored across layers for the first time. As we have shown in this review, there are already existing concepts of layer transitivity and paths but they do not explore geo-social interactions. In terms of extending multilayer scholarship with novel models and metrics, we introduce an interconnected geo-social network of people and places in the urban context where we extend the notion of structural holes and brokerage to places. Furthermore, we introduce the concept of serendipity which is unique to this type of models.

One of the biggest challenges in the field of empirical multilayer analysis has been the availability of data to demonstrate the utility and benefit of such models. This dissertation hopes to inspire novel ways of modelling existing data and the collection of datasets that take into account the multidimensional nature of human relationships in the geo-social context. Although narrow in applications at present, multilayer models are extremely useful in modelling complex systems realistically - including computer, urban and social systems - and with a future awareness of the presence of such tools and data, more interdisciplinary research around this will hopefully stem.

Chapter 3

Global Multiplexity

Timely statistics on key metrics of socio-economic status are essential for provision of services to societies, in particular marginalised populations. Despite the importance of accurate statistics to quantify the state of a country and progress towards favourable socio-economic outcomes, regular and reliable measurement is difficult and costly particularly in low income countries. With this in mind, in this chapter the network positions of countries in the global multiplex of international flows as well as individual networks is studied in order to approximate critical socioeconomic indicators of global importance. The techniques used favour simplicity and interpretability in order to be easily applied as potential benchmarks for UN efforts such as the Sustainable Development Goals (SDGs).

Following an introduction of multilayer models in Chapter 2, here we will demonstrate a concrete and novel application to the international development context. A simple but powerful multiplex approach to model the multiple interactions that exist between countries will be introduced in the first few sections of the chapter, where we will define the multiplex network as a collection of graphs. Furthermore, the *global degree*, which can be computed from the position of a node in the multiplex network is formalised and evaluated on a large multilayer dataset of six networks representing international interactions using correlation analysis. A *community multiplexity index* is also proposed to capture the connectedness of countries on a global scale and offer a simple and reliable way to estimate tie strength. The rest of the chapter will introduce the international postal network in detail, as well as the other five networks used to approximate critical socioeconomic indicators in the global multiplex. Finally, the potential of the community multiplexity index to approximate the socioeconomic profile of highly interconnected countries, even when challenged with missing data, is demonstrated through data-driven empirical evaluation.

3.1 Multiplex Model

As further elaborated upon in Chapter 2, a multiplex network is one where multiple connections exist between the same entities yet a different set of edges exists for a node in each layer [55]. In this dissertation, we will use a simple graph-based model of a *multiplex graph* $\mathcal{M}$, which can be defined as an ensemble of m graphs, each corresponding to an interaction type, and therefore representing a layer in the multiplex graph. The α-th layer of the multiplex is indicated as $G^{\alpha}(V^{\alpha}, E^{\alpha})$. Therefore, the collection of graphs composing the M-layer graph can be denoted as:

$$\mathcal{M} = \{G^1(V^1, E^1), ..., G^{\alpha}(V^{\alpha}, E^{\alpha}), ..., G^M(V^m, E^m)\} \tag{3.1}$$

An adjacency matrix A^{α} is associated with each graph $G^{\alpha}(V^{\alpha}, E^{\alpha})$ representing the layer α of the multiplex. Therefore, $\mathcal{M}$ can also be described with a sequence of adjacency matrices $A = [A^1, ..., A^{\alpha}, ..., A^M]$. We denote by a_{ij}^{α} the element of the matrix A^{α} at layer α representing the link between nodes i and j on that layer.

3.1.1 Global Multiplex Degree

Following from the definition of a multiplex model and similar approaches to multiplex degrees described in Chapter 2, we can also define the multiplex neighbourhood of a node i as the union of its neighbourhoods on each network layer:

$$N_{\mathcal{M}}(i) = \{N_{\alpha}(i) \bigcup N_{\beta}(i) ... \bigcup N_m(i)\} \tag{3.2}$$

where $N_{\alpha}(i)$ is the neighbourhood of nodes to which node i is connected on layer α.

The cardinality of this set can be considered as the node's global multiplex degree, or, in other words, the total number of neighbours with which a node has exchanges in any of the layers:

$$k^{glob}(i) = |N_{\mathcal{M}}(i)| \tag{3.3}$$

As described in more detail in Chapter 2, there are many possible combinations of layers to compute a node's degree in a multilayer network. In this initial analysis, we opt for the simplest version of a weighted degree which combines the layer weights of each edge linearly and under the assumption that they all add equally to the multiplex (although more analysis will be needed to assess this assumption). We can compute the weighted global degree of a node i as:

$$k_w^{glob}(i) = \frac{\sum_{j \epsilon N_{\mathcal{M}}(i)} \sum_{G \epsilon \mathcal{M}} e_{ji}}{n * m} \tag{3.4}$$

where for each neighbour in the multilayer neighbourhood $N_{\mathcal{M}}(i)$, we sum the number of of graph layers on which the edge appears and normalise by the total number of edges possible in the multiplex. We only consider bidirectional edges because the global degree is ultimately a measure of tie strength. This is common practice in contexts where tie strength is of importance such as in social networks [99]. We then normalise the weighted global degree by the number of possible edges $n * m$, where n is the total number of nodes and m is the number of networks in the multiplex collection.

3.1.2 Community Multiplexity Index

Networks are powerful representations of complex systems with a large degree of interdependence. However, in many such systems the network representing it naturally partitions into communities composed of nodes that share dependencies between each other, but share fewer with other components. We formalise this idea as the *community multiplexity index* of a pair of nodes (i, j):

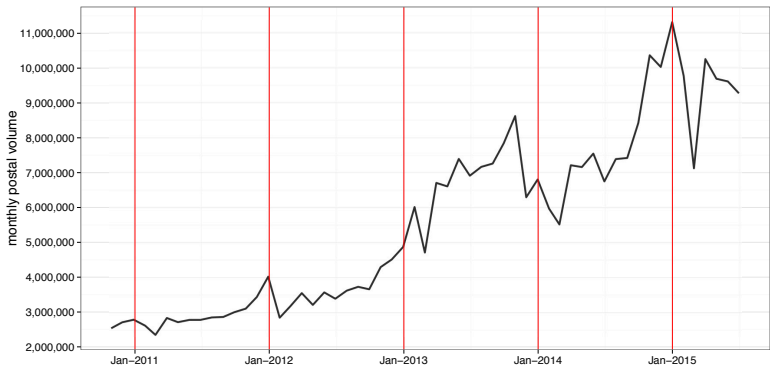
$$cmi(i, j) = \sum_{G \epsilon M} \delta(c_i^G, c_j^G) \tag{3.5}$$

where c_i, c_j are discrete variables indexing the clusters of which node i, j are members respectively. If the two are equivalent for a given network G, the level of community multiplexity increases by one, represented by the Kronecker delta function, which evaluates membership equivalency of the two nodes.

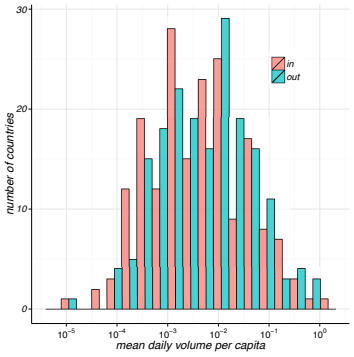
Prior work has explored information similarity in terms of community structure between layers [12, 90] and many novel ways of community detection in multilayer networks [52, 129, 147]. Although we use a community detection approach on each layer separately, our goal is not to obtain community clusters of countries in the multiplex but to observe the strength of connectivity between countries across layers as a measure of their similarity in order to build a proxy for exploring the socioeconomic similarity of pairs of countries.

3.2 Data

The network data used to demonstrate the above model and metrics consists of six layers of physical and digital interactions between 196 countries. Most of the data is openly available from the specified sources below apart from the international postal network, which is used as part of this thesis for the first time and described in further detail here.



(a) Global postal volume per month. The trend line indicates the seasonal peaks around the beginning of the year and the overall increase in postal volume over time.



(b) Average number of daily items sent (out) and received (in) per capita per country.

Figure 3.1: Postal volume

3.2.1 The International Postal Network

Although postal flows are understood to follow a gravity model [4], where flows between countries are dictated by volume and distance, similar to other networks describing flows, little is understood about the network properties of the postal network and how they relate to those of other global flow networks. The International Postal Network (IPN) is constructed using electronic data records of origin and destination for individual items sent between countries collected by the Universal Postal Union (UPU) since 2010 until present. Items are recorded on a daily basis amounting to nearly 14 million records of items sent between countries. As one of the most developed communication networks on a global scale, it is a dense network with 201 countries and autonomous areas, and 23,000 postal connections between them, with 64% of all possible postal connections established. The global volume of post has seasonal peaks observable in Fig. 3.1a. Notably, since 2010 postal activity is on the rise and this can be accounted for by the parallel growth of e-commerce [183]. This growth also positions postal flows as a sustainable indicator of socioeconomic activity.

In terms of daily activity, we can observe the mean relative number of daily items sent and received by countries during the period in Fig. 3.1b. This can be highly dependent on the size of the population of a country so these interactions have been normalised the volume per country's population. The annual population statistics provided by the World Bank and collected by the United Nations Population Division have been used for this purpose. From the distribution of volume, it becomes clear that the majority of countries send and receive a similar amount of post per capita, however, with a number of exceptions on both ends where a few countries send and receive exceptionally low or high number of items.

Next we can observe the degree distributions of both the weighted and unweighted global postal graphs shown in Fig. 3.2, as the complementary cumulative probability function

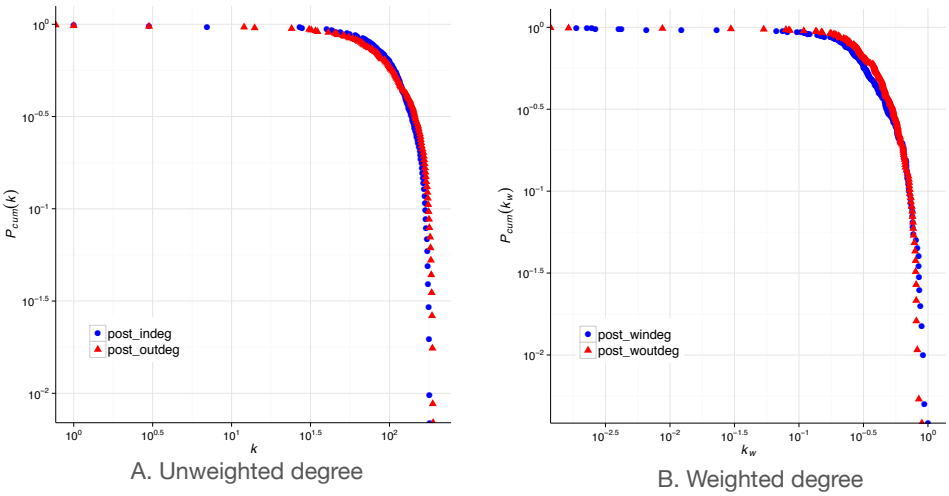


Figure 3.2: International Postal Network degree distributions

(CCDF). In Fig. 3.2a, we can see that the in and out degrees are relatively balanced in both instances and that about 50% of countries have more than 100 postal partners. The weighted degree in Fig. 3.2b follows a similar pattern but includes the weight of connections in the graph. The network is weighted by summing the total annual volumes of directed flow between two countries, averaged over years and normalised over the population of the country of origin. It is then further normalised by the maximum weight in the network, resulting in a value between 0 and 1, allowing for the comparison of values between networks. The weighted adjacency matrix of the top quartile of countries in terms of degree can be seen in Fig. 3.3 with the US and UK having the largest numbers of postal partners. Prominent postal network countries have relatively high interaction with most of their partners, including interactions with lower ranked countries. This is related to the degree assortativity (discussed in Section 2.2.4) within the postal network, elaborated upon in the following section.

3.2.2 Other global flow networks

This work builds upon previous efforts using global flow networks to present novel data sources for international development efforts and to demonstrate a holistic view of several distinct flow networks. We consider five networks, which have been previously studied independently, along with the IPN. We will now describe these networks and compare their properties in the following section.

The World Trade Network The trade network is constructed from records maintained by the UN Statistics Division in the Comtrade Database and provided by the MIT Atlas Project. It contains the number and value of products traded between countries classified by commodity class.

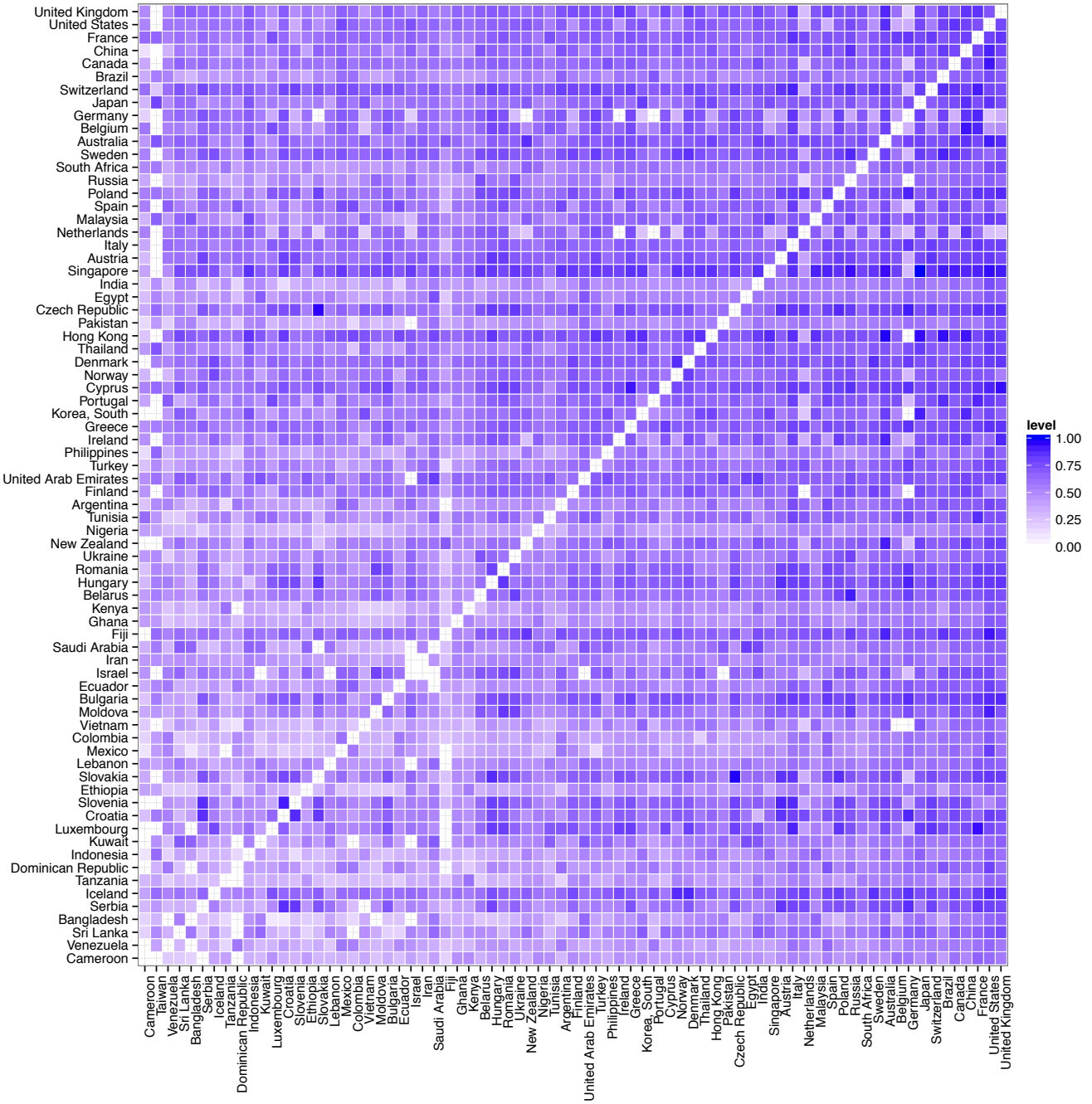


Figure 3.3: Matrix of the intensity of connections between countries based on the number of items exchanged (higher is darker); axes are ordered by the country's unweighted postal degree (its number of postal partners); only countries with more than 120 postal partners are displayed.

The Global Migration Network This is compiled from bilateral flows between 196 countries as estimated from sequential stock tables. It captures the number of people who changed their country of residence over a five-year period. This reflects *migration transitions* and not short term movements and is provided by the *Global Migration Project*.

The International Flights Network The flights data is collected by 191 national civil aviation administrations and compiled by the International Civil Aviation Organisation (ICAO). These tables detail, for all commercial passenger and freight flights, the country of origin and destination and the number of flights between them [81].

The IP Traceroute Network This city to city geocoded dataset is built from traceroutes in the form of directed IP to IP edges collected in a crowdsourced fashion by volunteers through the DIMES Project [167]. The project relies on data from volunteers who have installed the measurement software which collects origin, destination and number of IP level edges which were discovered daily. We aggregate this data on a country to country basis and use it to construct an undirected Internet topology network, weighted by the number of IPs discovered and normalised by population.

The Social Media Density Network Constructed from aggregated digital communication data from the Mesh of Civilizations project, where Twitter and Yahoo email data are combined to produce an openly available density measure of the strength of digital communication between nations [172]. This measure is normalised by the number of Internet users in each country and thus is well aligned with the other networks we use. It also blends data from two distinct sources and thus provides greater independence from service bias. We build an undirected network from this data where only bidirectional edges in the two platforms are considered.

Global Wellbeing Indicators In the following analysis we compare the above networks and use multiplexity theory to extract knowledge about the strength of connectivity across them. We distinguish between single layer and multiplex measures, which allows us to observe to a deeper extent the international relationships and the potential for using global flow networks to estimate the wellbeing of countries in terms of a number of socioeconomic indicators summarised in Table 3.1. These indicators are widely-used global benchmarks for the health and wellbeing of countries and have been compiled from a number of organisations and agencies concerned with the measurement of global conditions.

3.3 Comparing Multiple Networks

Although each of the five networks previously described, apart from the International Postal Network (IPN), has been studied separately, there has not been a comparative analysis of all. In Table 3.2, the network properties of all six networks are listed separately. The number of nodes or countries exceeds 195(6) due to differing lists of member states providing statistics to each authority. Although weights are distinct for each

Abbreviated	Full name	Description	Source
GDP	Gross Domestic Product	Aggregate measure of production on a on a per capita basis	The World Bank
LifeExp	Life Expectancy	Life expectancy since birth in years	The World Bank
CPI	Corruption Perception Index	Perceived levels of corruption, as determined by expert assessments and opinion surveys	Transparency International
Happiness	Happiness Score	Survey of the state of global happiness perceptions	Gallup World Poll
Gini.Idx	Gini Index	Income inequality on a national level	The World Bank
ECI	Economic Complexity Index	Holistic measure of the production characteristics of large economic systems	The Observatory of Economic Complexity
LitRate	Adult Literacy Rate	Percent of adult population who are literate	UNESCO
PovRate	Poverty Rate	Percent of population living bellow national poverty threshold	The World Bank
EdRate	Education Rate	Percent of population who have completed primary school	The World Bank
CO2	Emissions of carbon dioxide	Carbon dioxide in billions of metric tonnes per capita	Carbon Dioxide Information Analysis Center
FxPhone	Fixed Phone Rate	Percent of population living in households with a fixed phone line	Int Telecommunication Union
Inet	Internet penetration	Percent of population who have accessed the Internet in the past 12 months	Int Telecommunication Union
Mobile	Mobile cellular subscriptions	Percent of population who have a mobile cellular subscription	Int Telecommunication Union
HDI	Human Development Index	Composite statistic of life expectancy, education, and income per capita indicators	UNDP

Table 3.1: Description and source of the fourteen global indicators.

network	**weight**	**years**	**|V|**	**|E|**	**< k >**	**assort**	**d**	**cc**
Post	postal items	2010 – 15	201	22,280	110.85	-0.26	0.55	0.79
Trade	export value	2007 – 12	228	30,235	132.6	-0.39	0.58	0.84
Migration	migrants	2005 – 10	193	11,431	59.22	-0.33	0.31	0.68
Flights	flights	2010 – 15	223	6,425	28.81	-0.1	0.13	0.49
IP	IPs	2007 – 11	225	9,717	43.19	-0.42	0.19	0.6
SM	density	2009	147	10,667	145.13	-0.02	0.98	0.99

Table 3.2: Network Properties: number of nodes, number of edges, average (out) degree, degree assortativity, network density, average clustering coefficient

network, they always represent a volume of flow between areas. While there are small discrepancies between the years of each network, most networks cover a five year period, with the exception of the Social Media network which is from a single year. The volume of interaction between two countries is averaged over the number of years for each network.

All networks are weighted by normalising the raw volume of interaction described above by the population of each respective country of origin and rescaling all weights across networks within the same range [0,1] by dividing by the maximal weight, as we did for the

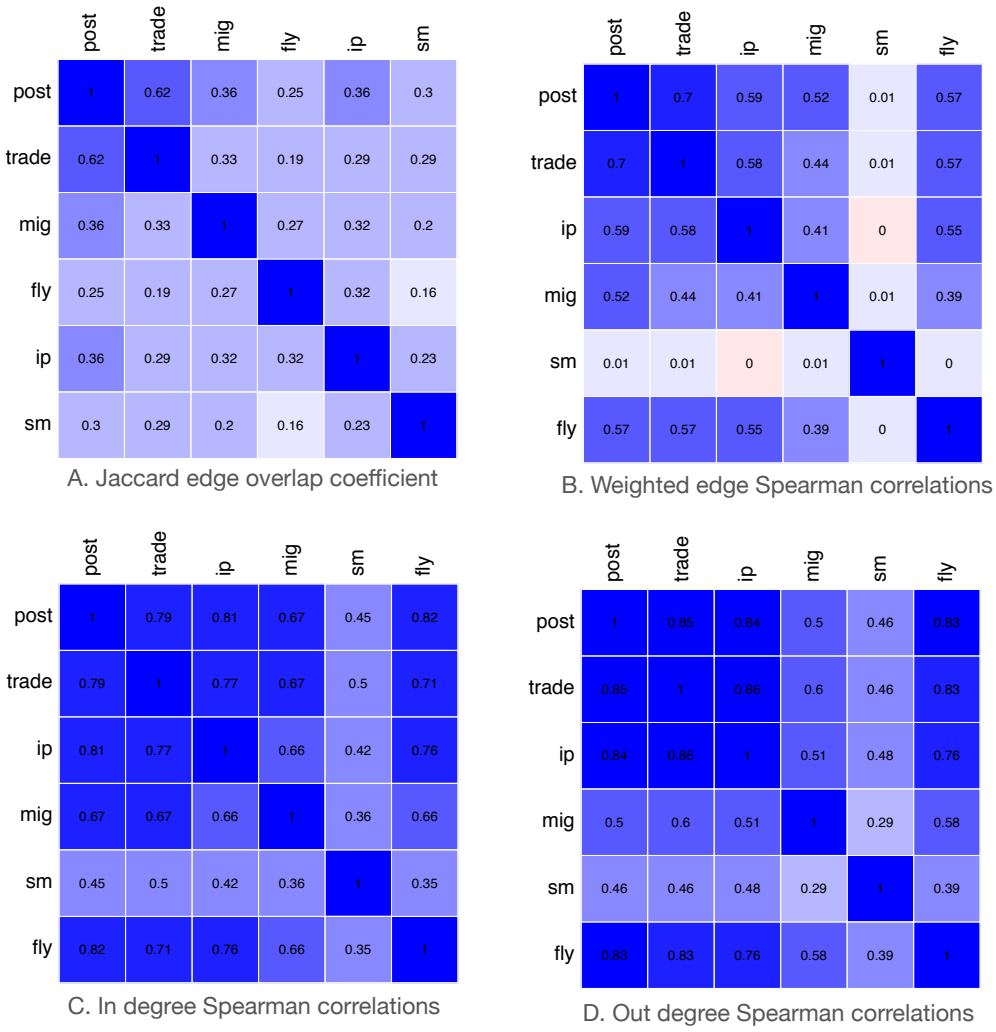


Figure 3.4: Comparative analysis of the IPN and other networks in terms of Jaccard overlap, percent shared edges, edge weight correlation and in and out degree correlations.

postal network in the previous section. The number of ip traceroutes originating in countries with higher population is higher in the city to city edge dataset from DIMES due to the crowdsourced nature of the project. Although, perhaps more appropriate normalisations may exist in an independent study, in order for the weights to be comparable across networks in the present research they are all normalised by population of the country.

The average out degree for each directed network is computed in a standard way as for the postal network, as well as the degree assortativity (Pearson correlation between the degrees of all pairs of connected countries), the network density and clustering coefficient. The assortativity coefficient determines to what extent nodes in the network have mixing patterns that are determined by their degree. Positive assortativity means that nodes with high degree tend to connect to other nodes with high degree, whereas a negative assortativity means that nodes with high degree tend to connect with others with lower degree, which is the case for all of the six networks as seen in Table 3.2. Although all networks differ in size and average degree, they have relatively high clustering coefficients,

reflecting a general tendency for countries to cluster together in global networks. This clustering however is not based on the importance of a node (its degree) since the assortativity coefficients for all networks are low or negative, suggesting that global networks are dissassortative and therefore higher degree nodes tend to connect to lower degree nodes.

Fig. 3.4 presents a comparative analysis between the six networks. They are referred to for short as: post, trade, ip, mig, sm and fly. The Jaccard coefficient, described as the number of edges that exist on both networks is divided over the number of edges that exist in any of the two networks, is used to compute the overlap of edges between pairs of networks in Fig. 3.4a. The highest Jaccard overlap is between the postal and trade networks, the two densest networks. The rest of the networks however are not strongly overlapping in terms of edges, which implies that each distinct network layer provides a non-trivial and complementary view of how countries connect. Nevertheless, the Spearman rank correlation between weighted edges in Fig. 3.4b reveals that the volume of flow of goods, people, and information is correlated for those edges between countries, which exist on both networks. A notable exception is the digital communications network (sm), which is entirely uncorrelated with any other network. This means that countries likely connect in unexpected ways on social media and email.

When considering the degree of a country as an indicator of its position in the network, we find that there are high correlations between the in and out positions of countries in Fig. 3.4c and Fig. 3.4d. Although to a lesser extent, the social media network is also correlated with the rest, despite its lower Jaccard overlap with other networks.

3.4 Approximating Indicators With Global Networks

In this section the networks discussed previously are compared to the values of the socioeconomic indicators in Table 3.1, with the goal of using network theory to approximate such difficult to obtain indicators. The network degree is computed for each of the six networks, defined as the sum of the neighbours for both incoming and outgoing connections where directed. This reflects how well connected a country is in a particular network. The weighted incoming and outgoing degrees on each network are also computed, defined as the sum of the normalised flows from all neighbours and reflecting the volume of incoming and outgoing flows. In addition to these standard single-layer network metrics, the previously defined *global degree* of a country, which takes into account connectivity across all networks is computed. A plot of the cumulative degree distribution of both the weighted and unweighted global degrees can be seen in Fig. 3.5.

The average global degree is 110 and the average global weighted degree is 250, which means that each country connects with an average of 110 other countries through two or more layers. In terms of unweighted degree (number of unique connections globally in the multiplex) in Fig. 3.5a, we notice a substantial curvature, indicative of the moderately

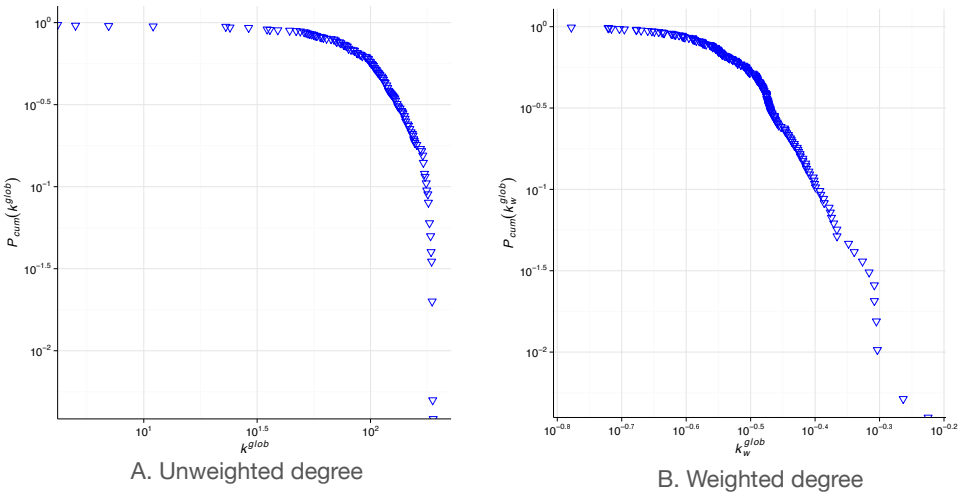


Figure 3.5: CCDF of weighted and unweighted global multiplex degrees.

stable degree approaching 10^2 but a sudden decline after, indicative of the few countries $10^{-0.5}(32\%)$ having a degree higher than 130. A steeper decline can be observed in the weighted distribution in Fig. 3.5b, where the majority of countries have a weighted degree of 0.25 or less ($10^{-0.6}$), signifying that they have realised 25% or less of their connectivity in the global multiplex. Although many empirical measurements of networks are noted to follow a power law distribution, this appears as a straight line in a log-log degree distribution plot, which is clearly not the case in the data. However, what is apparent is that the distribution is right-skewed, with a small number of countries being observed to have high global degrees.

Fig. 3.6 shows the Spearman rank correlation between the network degrees of the six networks (in and out degree, and weighted in and out degree) and various socio-economic indicators: GDP, Life expectancy, Corruption Perception Index (CPI), Internet penetration rate, Happiness index, Gini index, Economic Complexity Index (ECI), Literacy, Poverty, CO_2 emissions, Fixed phone line coverage, Mobile phone users, and the Human Development Index. These indicators and their significance for the international development agenda are described in detail in Table 3.1.

All degrees of single networks and the global degree appear vertically in Fig. 3.6 and all indicators appear horizontally. In general, weighted outgoing degrees on the single networks perform best for the post, trade, ip and flight networks. An exception from the physical flow networks is the migration network, where the incoming migration degree is more correlated with the various indicators. The best-performing degree, in terms of consistently high performance across indicators is the global degree (for 7 out of all indicators). *This suggests that looking at how well connected a country is in the global multiplex can be more indicative of its socioeconomic profile as a whole than looking at single networks.* The weighted global degree performs slightly worse in the Spearman ranked correlations which emphasises the importance of structural properties of connec-

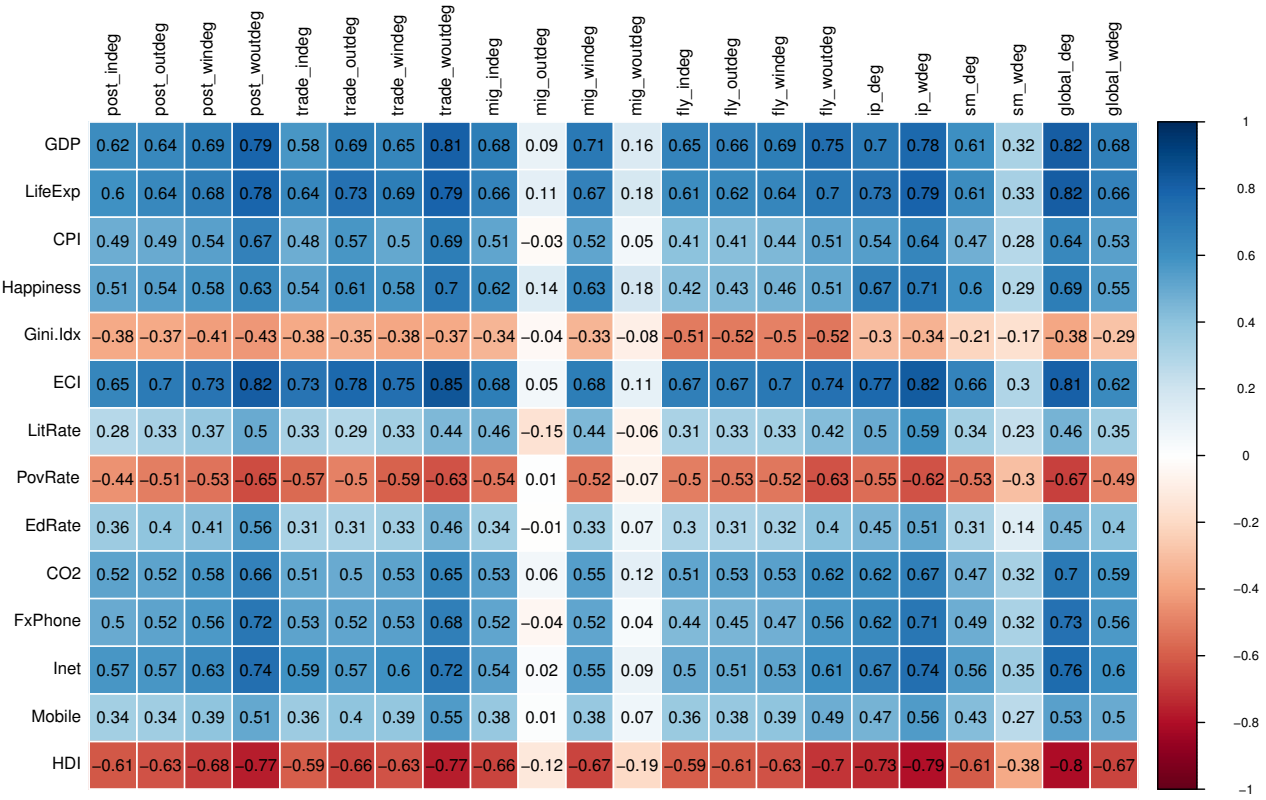


Figure 3.6: Spearman rank correlations between global flow network degrees and socioeconomic indicators.

tivity rather than aggregate annual flow volumes when it comes to international country rankings. For example, a country may have few intense flow volume links but be overall rather disconnected in the network. Although as an initial study the simplest degree measures in multiplex networks as described in Chapter 2 were used here, future work would benefit from a more fine-grained temporal analysis of weighted multiplex degree measures where such data is available.

The GDP per capita and life expectancy are most correlated with the global degree, closely followed by the postal, trade and ip weighed degrees. This shows a relationship between national wealth and the flow of goods and information. The perception of corruption index (CPI) however, is most positively correlated with the out weighted degrees of the postal and trade networks, followed by the ip network, similar to their relationship with the happiness index. This signifies that less corrupt and happier countries have greater outflows in those respects. On the other hand, the Gini Index of inequality is distinctly most negatively correlated with the flight network, which means that countries with greater inequality have less incoming and outgoing flight connections. The ECI index is equally highly correlated with most network degrees, and especially the global degree, trade, ip and post degrees. Literacy, Education and mobile phone users per capita were more weakly correlated across than other indicators, which means that there may be better predictor variables beyond the scope of this work for those indicators. Fixed

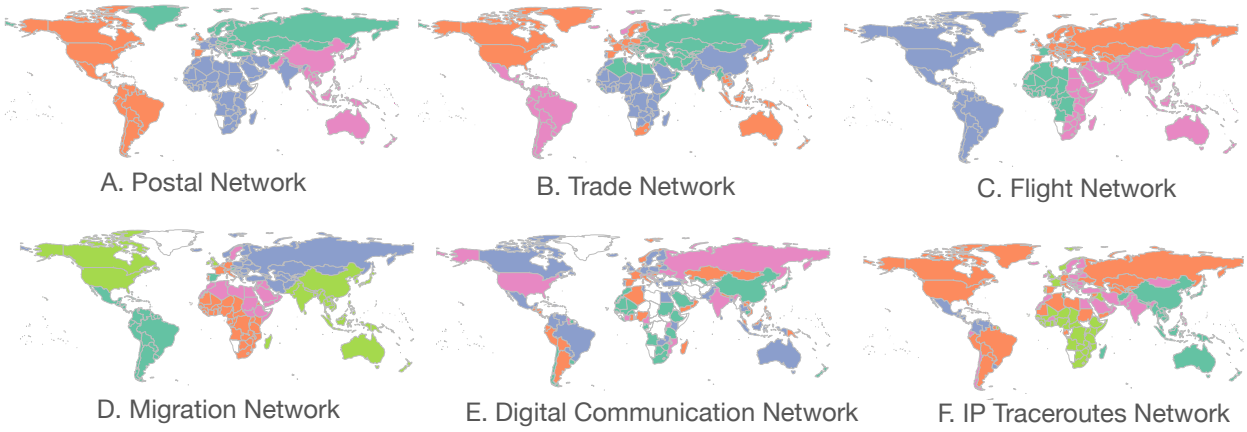


Figure 3.7: Country community membership for each network.

phone line households, Internet penetration and CO_2 emissions, however, are positively correlated with the global degree, followed by the postal and ip degrees. This indicates the importance of global connectivity across networks with respect to these factors.

Similarly to GDP, the rate of poverty of a country is best represented by the global degree, followed by the postal degree. The negative correlation indicates that the more impoverished a country is, the less well connected it is to the rest of the world. Finally, one of the most strongly correlated indicators with the various degrees is the Human Development Index (HDI), low human development (high rank) is most highly negatively correlated with the global degree, followed by the postal, trade and ip degrees. This shows that high human development (low rank) is associated with high global connectivity and activity in terms of incoming and outgoing flows of information and goods.

3.5 Global Community Multiplexity Index

In the previous section, network measures were related to various socioeconomic indicators, showing that metrics such as the network degree can be used to estimate wellbeing at a national level. In this section, we further examine the connectedness between pairs of countries through community structure across network layers as a form of socioeconomic similarity. In the present context, communities are composed of groups of countries that share higher connectivity than the rest of the network. If two countries appear in the same community across many network layers, this can be considered a greater level of connectivity and an indicator of greater socioeconomic similarity, otherwise not visible from the single network perspective. We use the Louvain modularity optimisation method [20] for community detection in each individual network (as described in Chapter 2), which takes into account the tie strength of relationships between countries and finds the optimal split in terms of disconnectedness in the international network. This returns between 4-6 communities for each network, the geographical distribution of which is shown in Fig. 3.7.

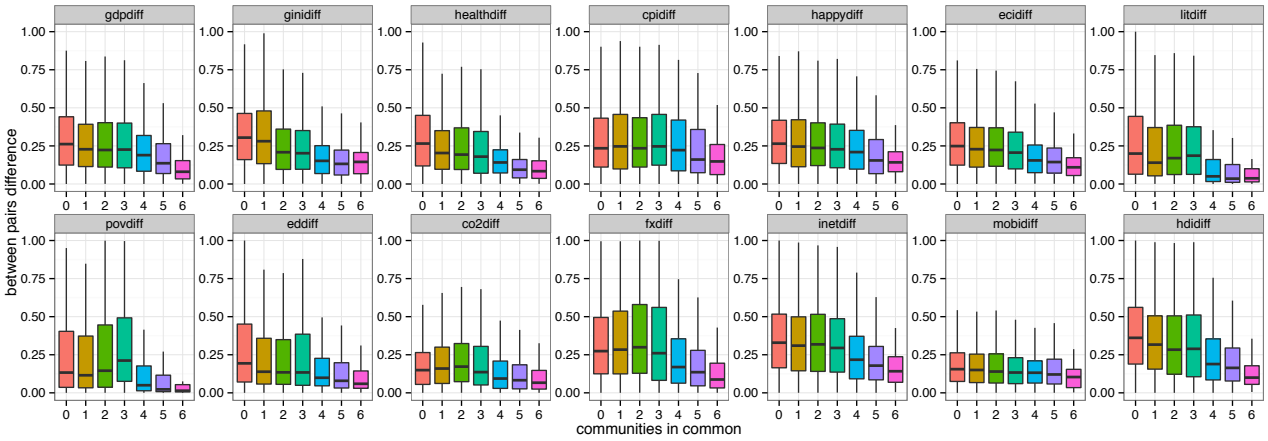


Figure 3.8: Socioeconomic difference margin between countries who share communities in the global flow networks.

Although communities naturally seem to be very driven by geography in physical flow networks, this is not the case in digital networks where communities are geographically dispersed. This is an indication of the difference in the way countries connect through post, trade, migration and flights rather than on the IP and social media networks. However, *what does it mean for two countries to be both members of the same network community?* Common community membership indicates a level of connectedness between two countries, which is beyond the randomly expected for the network. It is often observed that nodes in the same communities share many similar properties, therefore it can be expected that pairs of nodes which share multiple communities across networks are even more similar. Here, the overlap in pairwise membership between pairs of countries across the six networks is measured as the *community multiplexity index*, a measure of socioeconomic similarity.

The hypothesis is that countries that are paired together in communities across more networks are more likely to be socioeconomically similar. Similarity is measured here as the absolute difference between each indicator from the previous section for two countries and plot that against their community multiplexity. For example, the United States has an average life expectancy of 70 years, whereas Afghanistan has an average life expectancy of 50, the absolute difference between the two is 20 which represents low similarity when compared to the United Kingdom's life expectancy of 72 for this indicator.

In Fig. 3.8, we can observe the variations in similarity for countries with different levels of community multiplexity. What is immediately striking is that countries that share a maximal number of communities and therefore exhibit the greatest community multiplexity, have the smallest margin of difference across all indicators. This suggests that *countries with high community multiplexity index have a more similar socioeconomic profile than those with low community multiplexity index.* This is confirmed by a two-sample Kolmogorov-Smirnov (KS) test between the distributions of differences in each indicator

for pairs sharing different numbers of communities. Although the KS statistic is lower between groups sharing 0 and 1 communities (apx. 0.1 for all indicators and p-value <0.01), it is very high for groups between 1 and 6 communities (0.4 and above, p-value <0.01), except for mobile phone penetration (detailed KS test results are presented in Table 3.3).

Indicator	**KS test**	**p-value**
GDP	0.44	4.441e-16
Gini Index	0.39	1.068e-12
Health (LifeExp)	0.4	6.42e-13
CPI	0.25	1.759e-05
Happiness	0.34	1.107e-09
ECI	0.37	4.151e-11
LitRate	0.33	1.015e-06
PovRate	0.46	5.762e-07
EdRate	0.27	1.013e-06
CO2	0.29	1.566e-08
FxPhone	0.4	3.896e-13
Inet	0.38	4.463e-14
Mobile	0.25	2.693e-06
HDI	0.48	2.2e-16

Table 3.3: Two-sample Kolmogorov-Smirnov test statistic results and p-values for socioeconomic indicator differences between pairs of countries with minimal and maximal community multiplexity values (1 and 6).

Further to this observation, in most indicators there is a very strong significance in the level of community multiplexity - *the higher the community multiplexity index between two countries, the smaller the difference between their socioeconomic profiles.* There are notable exceptions to this such as the mobile phone penetration ratio, where it appears that beyond the highest level of multiplexity, all other countries are relatively similar in this aspect with low variation even for those pairs of countries which share no communities. For all other indicators such as GDP, Literacy ratio, HDI and Internet penetration, there is a dramatic increase in similarity past a community multiplexity of 3. Ultimately, these similarities can be used to estimate the wellbeing of countries for which it is unknown but can be estimated from its neighbours as suggested in this section but further analysis and predictive modelling will be required to support this.

3.6 Related Work

Previous work has explored flows of both physical and digital nature, where physical flows of goods and people [13, 1, 166, 85, 81, 112] and digital flows of information and communication [157, 63, 182, 120, 172] have been extensively studied in the past in order to understand better the way in which they affect the wealth, resilience and function

of social systems on global, regional, national and sub-national scales. More recently, these same data sources and methodologies have begun to be used to assist humanitarian and development organisations, allowing new ways to use data to implement, monitor and evaluate programs and policies [149]. The ability of such novel data sources to complement traditional data collection techniques such as household surveys and focus groups is clear [130]. The data is collected passively without the need for costly and potentially dangerous active data collection, which also avoids inaccuracies due to human error, bias [5] or dishonesty.

The use of data for development is still relatively nascent and questions remain over the ability of such sources to measure or approximate metrics of interest. Invariably, data sources such as social networking applications enjoy deeper penetration in developed economies and rely on expensive technologies such as smart phones and robust communications infrastructure. It has been noted that measurements of human dynamics based on such recent platforms can lead to strong biases [181], with worse implications for those with limited access to these digital platforms. Nevertheless, there are a few notable uses of urban transport data and CDRs for development applications. On the urban level, where more detailed records of movement are available, deprivation can be predicted from the flows of people through its transit system [168], the topic they tweet about, and its sentiment [153, 151]. Furthermore, evidence of social segregation can be found from such data, where it has been noted that more deprived areas receive inflows of more diverse origins than less deprived areas [101].

Closer to the international scope of this work, the authors in [169] explore the potential of aggregated CDRs to predict poverty in countries where other sources of data might be scarce at the cell tower level. Previous work has also attempted to use machine learning methods on the city-level with similar data [171] and the authors in [121] explored how top-up amounts reflect on poverty indicators. These methods however, have not been demonstrated on a global level and would not be possible for every country due to proprietary CDRs. Other creative ways of monitoring poverty have been proposed using Night Time Light (NTL) measured from satellite imagery, where the levels of light have been correlated with GDP [66]. This method is very coarse grained where sometimes individual countries are difficult to tell apart and it might work for some regions better than others. In this chapter, we have demonstrated an approach using open data from international agencies and digital media to predict a number of crucial socioeconomic indicators, from subjective wellbeing to inequality. Most importantly, we have shown that a combination of these under the multiplex paradigm can provide a number of useful proxies and reliable metrics for monitoring global wellbeing.

3.7 Conclusions

The digital exhaust left by flows of physical and digital commodities provides a rich measure of the nature, strength and significance of relationships between countries in the global network. This chapter examined how these traces and their multilayer network structure can reveal the socioeconomic profile of different countries. By measuring the position of each country in the Trade, Postal, Migration, International Flights, IP and Digital Communications networks, the potential to build proxies for a number of crucial socioeconomic indicators such as GDP per capita and the Human Development Index ranking along with twelve other indicators used as benchmarks of national well-being by the United Nations and other international organisations was explored. In this context, a global connectivity degree measure was proposed and evaluated, showing the utility of multilayer analysis in light of international development challenges such as sustainability and missing data.

We have observed how the network properties of global flows can approximate critical socioeconomic indicators and how network communities formed across physical and digital flow networks can reveal socioeconomic similarities. Real-time measurements of international flow networks can ultimately act as global monitors of wellbeing with positive implications for international development efforts. Using knowledge about the way in which countries interact through flows of goods, people and information, we can use the principles of multiplexity theory to understand the strength of international ties and the network communities they form.

Although these results do not provide insight into the cause of the socioeconomic circumstances of a country, one explanation is that network measures derived from global flow networks are a proxy of socioeconomic activities and therefore highly correlated with the explored indicators. It is an open question as to whether a highly central position in the network leads to favourable socio-economic outcomes or vica-versa. The structural connectedness of a country in the global network represents the number of opportunities a country has to exchange goods, information and resources with our countries - the more opportunities, the higher the exchange and therefore socioeconomic benefit. The following chapter further explores the theory of multiplexity and its application to multilayer interpersonal networks at small scale and online social networks at a larger scale.

Chapter 4

Multiplexity and Tie Strength

In the previous chapter we observed how a simple multiplex framework can successfully model international relations across layers of physical and digital interactions. In this chapter we will further explore two types of multiplex aggregations, which will help us formalise a multiplex weight for tie strength in social networks. We evaluate the measure on the MIT Social Evolution dataset, where we will observe how different configurations of the three layer multiplex of interactions between students can capture different relationships between them. In this way, we validate the theory of media multiplexity, which states that pairs of people who communicate through a greater number of channels have a closer relationship than those who communicate through a few.

This chapter takes two complementary approaches to tie strength, that of computational social science and machine learning. While in the first set of experiments simple set theory and probability shows the relationship between multiplexity and tie strength, the second part leverages this to predict links in online social networks yet using the same network model. By examining a number of classic link prediction features in the multilayer context of two heterogeneous platforms - Twitter and Foursquare - and by introducing two multilayer features, we predict links across the two as opposed to training and testing on the same network. A framework of multilayer link prediction is defined and evaluated on a two-layer dataset of Foursquare and Twitter across three cities with AUC scores of up to 0.86. The multilayer features we introduce are shown to have greater predictive power than single-layer features in the online social network "ecosystem". Overall, this chapter demonstrates the applications of generalisable multilayer network models using different techniques applied to the online and offline social network context.

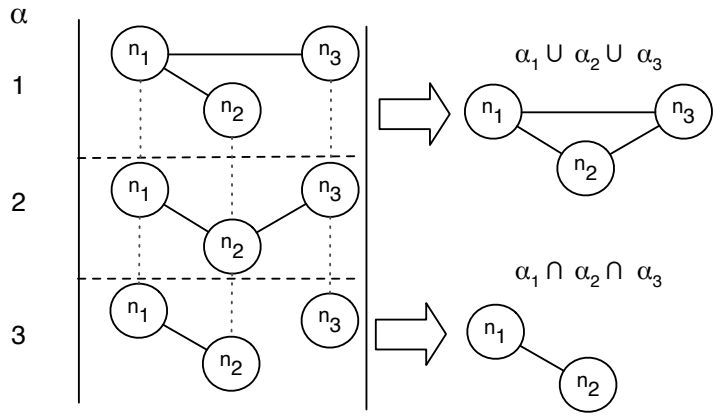


Figure 4.1: Union and Intersection aggregations are shown on the right, where α indicates the layer; the original networks/layers of the multiplex are shown on the left.

4.1 Multiplexity and Aggregations

As elaborated upon in Chapter 2, graph aggregations in multilayer networks have been a popular approach to multilayer network analysis. In the context of social networks, where each layer in the multiplex represents a different type of interaction between actors, various aggregate configurations can reveal novel information about the strength and *type* of their relationships.

Here, we build on the previously introduced graph-based model in Equation 3.1 and consider two distinct configurations. The Union aggregation, which takes the union of all edges in the multiplex represents the full multiplex connectivity between actors and can be useful in problems such as information diffusion. The union is defined as $G^{\alpha} \bigcup G^{\beta}$ of two graphs G^{α} and G^{β} represented respectively by the adjacency matrices A^{α} and A^{β} as the graph described by the adjacency matrix $A^{\alpha \cup \beta}$ with elements $a^{\alpha \cup \beta} = 1$ if $a^{\alpha}_{ij} = 1$ or $a^{\beta}_{ij} = 1$, 0 otherwise.

On the other hand, the Intersection aggregation, which takes the intersection of all layers in the multiplex. This perspective can be useful in considering problems related to resilience and reliability. The intersection is defined as $G^{\alpha} \bigcap G^{\beta}$ of two graphs G^{α} and G^{β} represented respectively by the adjacency matrices A^{α} and A^{β} as the graph described by the adjacency matrix $A^{\alpha \cap \beta}$ with elements $a^{\alpha \cap \beta} = 1$ if $a^{\alpha}_{ij} = 1$ and $a^{\beta}_{ij} = 1$, 0 otherwise.

Given these definitions, the following two types of layer aggregations of the multiplex $\mathcal{M}$ can be derived: *the union graph* $\bigcup_{\mathcal{M}}$ defined as

$$\bigcup_{\mathcal{M}} = G^1 \bigcup G^2 \ldots \bigcup G^M \tag{4.1}$$

i.e., the graph aggregation in which an edge between two nodes is present if it is present in *at least* one layer; and the *intersection graph* $\bigcap_{\mathcal{M}}$ defined as:

$$\bigcap_{\mathcal{M}} = G^1 \bigcap G^2 \ldots \bigcap G^M \tag{4.2}$$

i.e., the graph aggregation in which an edge between two nodes is present if it is present in *all* the layers. Both the union and intersection aggregations are illustrated in Fig. 4.1. The union aggregation is a graph configuration of the global connectivity and therefore, the multilayer neighbourhood $N_{\mathcal{M}}(i)$ can be used to describe the local ego-network of a node as per Equation 3.2.

Similarly, the *core neighbourhood* of a node i can be defined in the intersection configuration as:

$$N_{\mathcal{C}}(i) = \{j \in V^{\mathcal{M}} : e_{i,j} \in E^{\alpha \cap \beta}\} \tag{4.3}$$

where the set of multiplex links is defined as $E^{\alpha \cap \beta}$. We can further consider the set of all single-layer links on layer α only as $E^{\alpha \setminus \beta}$. It is also worth noting that it is possible to restrict this aggregation to a subset of graphs $\{\alpha, \beta, \gamma, ...\}$ and define, for example, the union graph over the set of layers $\{\alpha, \beta, \gamma, \ldots\}$ as the graph $\bigcup_{\mathcal{M},\{\alpha,\beta,\gamma,...\}}$ corresponding to the union of the graphs of layers $\alpha, \beta, \gamma, \ldots$. The intersection graph aggregation over a set of layers can be defined in a similar way.

4.1.1 Data

The open-access MIT Social Evolution dataset contains details of the everyday life of a group of students between October 2008 and May 2009 [117]. These students co-reside in two adjacent college residence buildings during term time. Details of their health habits, political orientation, music preferences, social relationships (online and offline), and mobile communication were collected during this period, allowing for a rich analysis of the relationships between their characteristics, social ties and communication.

Communication Layers The multiplex interaction graph is built by combining different communication layers. Three types of interactions can be extracted from the mobile phone data - physical proximity data (whether pairs of users were within 10 meters of each other, inferred from Bluetooth); phone call record data (who called whom); and SMS data (who texted whom). Each of these communication layers is a network in its own right (Table 4.1).

The degree distributions show that each communication layer is utilised to a different extent and purpose (Fig. 4.2). While the proximity layer has a high average degree, the other two layers have a low average degree in comparison, giving us initial insight into the social dynamics of the student community - many students meet many others but

network	type	avg degree	nodes	edges
calls	directed	5.8	69	401
SMS	directed	2.12	33	70
proximity	undirected	61.2	74	4,526

Table 4.1: Network specifications. *We take into account only interactions between students, ignoring external ones in the dataset.*

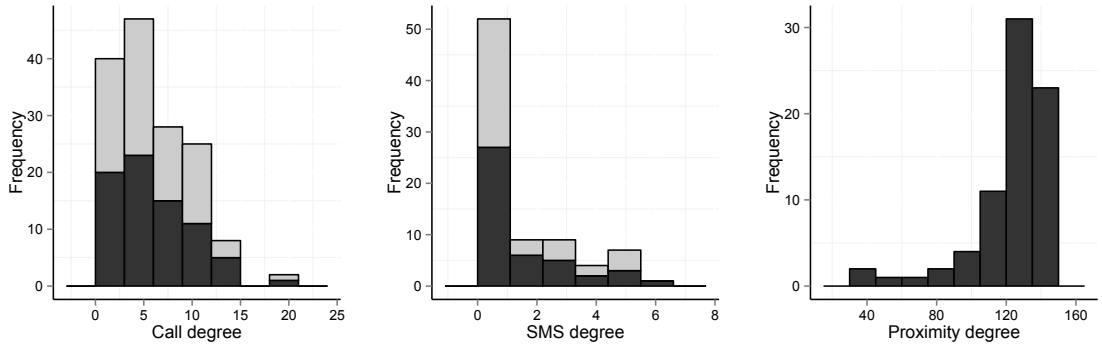


Figure 4.2: Degree distributions for the call, SMS and proximity networks. In-degree is in light, while out-degree is in dark (proximity is undirected).

few talk to many on the phone or text. The three layers complement each other, and in combination represent three basic communication channels of human interactions - spending time together, talking on the phone and sending messages.

If we denote the set of edges in the proximity layer as P, the call layer as C and the SMS layer as S, the relationship between the three communication layers can be described as $S \subset C \subset P$, meaning that all participants have been co-located with another participant and are part of the proximity layer but not all have called or sent a text message to another participant. This is because the proximity layer is prevalent, likely due to the fact that all students are co-residing and possibly have lectures together. All pairs with a call edge or a SMS edge also have a proximity edge. Incidentally, almost all pairs with a SMS edge also have a call edge (92% overlap), which may not be generalisable to the case of other communication networks. Overall, the density of the student network is such that 83% of all nodes are connected on at least one layer - the proximity layer.

Social Relationships Social relationships reported by the participants form the ground truth for the social tie analysis. Details of data collection methodology are described in [117]. Three types of reported social relationships are considered: *Facebook friendship*, *Socialising twice per week*, and *Close friendship*. The relationships are not mutually exclusive. We consider a pair to have a given relationship if i has declared that relationship with j in at least half of the six surveys during that time period (reports were given approximately every month and a half). We allow these relationships to be directed and not just reciprocal. For example, if i calls, sends messages to and meets with j

category	parameters	range	summary	*avg*	*SD*
political	interest in politics	0-3	t_{max}	1.67	1.00
	political orientation	1-7	t_{max}	5.34	1.31
health	weight(lb)	min 81.00 - max 330.00	t_{avg}	157.5	41.34
	height(in)	min 60.00 - max 81.00	t_{avg}	67.4	4.14
	salads per week	min 0.00 - max 6.00	t_{avg}	1.46	1.43
	fruits per day	min 0.00 - max 7.00	t_{avg}	2.12	1.45
	aerobics per week(days)	min 0.00 - max 7.00	t_{avg}	1.91	1.9
	sports per week(days)	min 0.00 - max 6.00	t_{avg}	0.89	1.5
music	indie/alternative rock	0-3	t_{max}	1.75	1.17
	techno/lounge/electronic	0-3	t_{max}	1.34	1.09
	heavy metal/hardcore	0-3	t_{max}	1.01	1.1
	classic rock	0-3	t_{max}	1.84	1.1
	pop/top 40	0-3	t_{max}	1.23	1.08
	hip-hop r&b	0-3	t_{max}	0.75	0.86
	jazz	0-3	t_{max}	1.19	1.03
	classical	0-3	t_{max}	1.76	1.09
	country/folk	0-3	t_{max}	0.84	0.96
	showtunes	0-3	t_{max}	1.25	1.14
	other	0-3	t_{max}	1.25	1.23
situational	year in college	1-5	*actual*	2.5	1.37
	residential sector	1-8	*actual*	4.9	2.16

Table 4.2: Survey parameters and summary over time. Three types of summary are considered over time: t_{max} when the final reported value is taken in the final survey of the study, t_{avg} is the average of all reported values and *actual* is when the actual value is reported.

(full connectivity in the multiplex), and considers him a close friend (maximal social relationship), however j does not reciprocate the relationship or communication (minimal connectivity and social relationship reported), j is still considered a close friend of i according to our definition because i treats him as such.

The set of reported close friendships edges can be denoted as CF, the set of those who reported socialising as SC, and the set of Facebook friendships as FB, then the relationship between the three can be described as $CF \subset SC \subset FB$. This signifies that all close friends socialise and are Facebook friends, but not all Facebook friends socialise and are close friends.We assign the highest subset (most inclusive) set to a pair, so pairs have a single definitive social relationship for the purpose of our analysis. Overall, we have 2,179 directed pairs who have not declared any social relationship; 1,299 who are only friends on Facebook but do not socialise regularly; 586 who do socialise twice per week but are not close friends, and 462 pairs of close friends. If we split relationships according to online and offline presence, we can state that all social relationships have online presence in the form of Facebook friendship (all 2,347), while all non-declared social relationships do not (all 2,179).

User Profiles From survey data collected periodically over the study period, there is additional information about the participants - health, political and music preferences, as well as their residential sector and year of study. We summarise the information from each category to create a composite view of a participant's attitude. Table 4.2 contains a description of this data, further elaborated upon next.

- **Political.** Information about the participants' political sentiments around the 2008 presidential election. This information includes the participant's level of political interest ranging from *Very interested - 3*, *Somewhat interested - 2*, and *Slightly interested - 1* to *Not at all interested - 0*; and political orientation from *Extremely liberal - 7*, *Liberal - 6*, *Slightly liberal - 5*, *Moderate middle of the road - 4*, *Slightly conservative - 3*, and *Conservative - 2* to *Extremely conservative - 1*. The liberal to conservative scale is fairly fine grained and could be independent from a political party. On average, participants reported they are *Slightly liberal* to *Liberal*. Since the political orientation can evolve over time and may be affected by the election period as was reported in [118], the value of these parameters is reported at the end of the observation period for all students, denoted by t_{max}.

- **Health.** The average of the weight, height, salads and fruits, and aerobics and sports per week allowed us to build a comprehensive health profile for each user. An average (t_{avg}) over the survey period is used to give a single value for each attribute, and gain an understanding of the overall health habits of each student. Previous studies on this dataset found that tie formation was strongly dependent on health factors such as aerobic exercise and other campus activities [59].

- **Music.** We use the self-reported interest in each genre to build a music profile for each student. There are 11 different genres to which users have attached a preference ranging between *No interest - 0*, *Slight interest - 1*, *Moderate interest - 2* to *High interest - 3*. The most popular genre is "classic rock" with an average rating of 1.84 and the least popular one is "hip-hop and r&b" with an average rating of 0.75. The association between homophily and music has been notably drawn in [122], where music types were found to create niches in socio-demographic segments of society.

- **Situational.** We have information about the residential sector (floor and building, where there are two buildings separated by a firewall only) and the college year of each student during the academic year. We coded each sector from 1 to 8 according to location and adjacency to each other. For example, sectors 1 and 2 are on the same floor, separated by a firewall, sector 3 is in the same building as sector 4 but one floor up, while sector 5 is adjacent to it and so on. In terms of year in college, students are either a *Freshman - 1*, *Sophomore - 2*, *Junior - 3*, *Senior - 4* or a *Graduate Resident Tutor - 5*. These situational factors have been previously found to be highly indicative of a social tie [59].

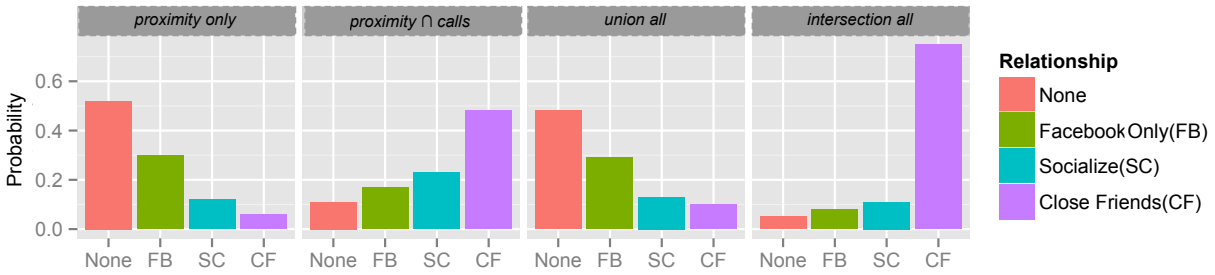


Figure 4.3: Probability mass function $P(X = x)$ for each declared relationship x within each graph aggregation. *None (None)* means there was no declared relationship in the directed pair; *Facebook Only (FB)* means that there was only a Facebook friendship declared between the pair; *Socialise (SC)* means that the pair declared they socialise twice per week in addition to being friends on Facebook; and *Close Friends (CF)* means that there was a declared close friendship in addition to all other social relationships.

4.1.2 Multiplex Social Ties

In media multiplexity, the use of many different media as a means of communication indicates a strong tie [83]. Here, the strength of online and offline ties will be examined by considering the multiplexity of communication between students. Due to the relationships between the layers (denoted as S: SMS, C: calls, and P: proximity) being $S \subset C \subset P$, there exist four possible non-redundant aggregations: (1) proximity layer only, which includes edges present *only* on the proximity layer and no other layer; (2) $proximity \bigcap calls$, where we have those edges present on both the proximity and call layer (equivalent to set C); (3) $proximity \bigcup calls \bigcup SMS$, or the union of all layers (all pairs on any layer); and (4) $proximity \bigcap calls \bigcap SMS$, or the intersection of all layers (only pairs on all layers, equivalent to set S). Each of these configurations represent a different set of pairs according to their structural overlap in the multiplex network.

In Fig. 4.3, the probability of each of the three social relationships (and that of having no relationship) are compared within each of the aggregations described above. For each one, the probability of having a given social tie is measured as the probability density function (PDF) of a variable. Overall, it can be observed that single-layer communication is indicative of no relationship or an online social tie. As the number of layers increases to two, the probability of having no tie decreases dramatically, while the probability of having a stronger offline tie increases. At the highest level of multiplexity, which in this case is three layers, the highest probability of friendship exists. This is aligned with previous studies of media multiplexity, and demonstrates the same principle with just a few mobile communication layers.

Most strikingly, the highest probability of close friendship ($P = 0.75$) occurs at the intersection of the three layers (*intersection all* in Fig. 4.3). In this aggregation, all other social ties are underrepresented, highlighting the relationship between high multiplexity and strong ties (close friendship). The union and proximity only aggregations reflect the probability of having no social relationship ($P = 0.5$), and also of being Facebook friends

only ($P = 0.3$). The intersection between the proximity and call layers on the other hand ($proximity \bigcap calls$), gives a more balanced representation of the different relationships with a 0.5 probability of close friendship, a 0.23 probability of socialising twice per week, and a 0.17 probability of being friends only on Facebook. The total probability of being friends on Facebook if two students have met during the period is defined by the total probability of a social tie, since all social ties (all relationships except "None") are also present on Facebook. This gives a 0.5 probability of being friends online if the pair is connected on one layer (in our context the proximity layer), 0.9 if connected on two (proximity and calls), and with certainty if connected on all layers.

Given the above observations, it can be concluded that *the more communication channels utilised, the stronger the tie*, and that the level of multiplexity is a good indicator of tie strength. If the number of layers is considered as an indicator of tie strength, the strength of a tie can be described in terms of a multiplex edge weight in the network as:

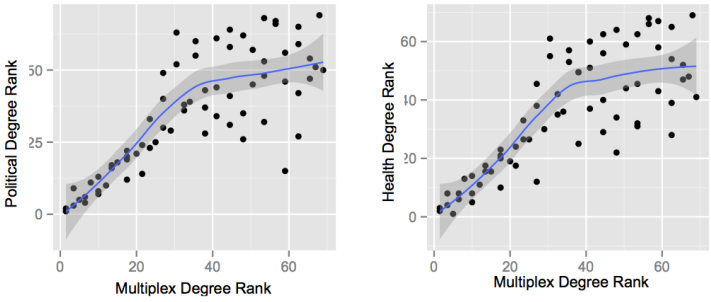
$$mw_{ij} = \sum_{\alpha=1}^{M} \frac{a_{ij}^{\alpha}}{M} \tag{4.4}$$

where M is the total number of layers in the multiplex, α is the layer in the multiplex as per Equation 3.1 and a is the edge weight in the adjacency matrix A. For example, if two students (i and j) utilise all possible channels for communication (in this case $M = 3$), mw_{ij} will be equal to 1, whereas if they use one channel, mw_{ij} will be 1/3 .

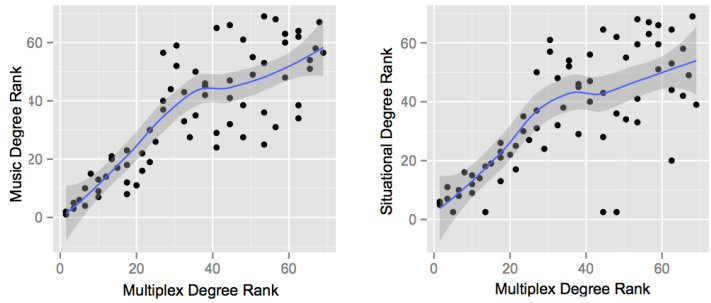
4.1.3 Homophily and Multiplexity

In this section, the homogeneity and homophily in the community as it relates to multiplex tie strength will be examined where politics, music, health habits, residential sector and year in college will be considered as diversity factors. *Diversity* in this context defines the variety of information introduced by ties as opposed to *homogeneity*. Based on Fischer's observation that greater multiplexity results in greater similarity [70], it can be expected that on both the edge and neighbourhood network levels, *the more communication channels utilised, the greater the similarity observed with respect to politics, health, music and residence & year in college.*

Profile Similarity & Multiplexity With the presumption that individuals with stronger (multiplex) ties bear greater similarity, the similarity between student profiles are compared to the strength of their multiplex relationship as defined by the multiplex weight (mw). The similarity scores between students are derived using the cosine similarity of the vector of attributes for each category - music, health, political, and situational for each pair of students. These values are described in detail in Table 4.2. As an example, if two students are both *Somewhat interested* in politics ($value = 2$), and one is *Slightly liberal*



(a) political (b) health



(c) music (d) situational (year/floor)

Figure 4.4: Spearman degree rank correlations (ρ) between similarity and multiplex network. (a) political $\rho = 0.78$ (b) health $\rho = 0.79$ (c) music $\rho = 0.81$ (d) situational $\rho = 0.73$, all p-values < 0.01.

($value = 5$), while the other is *Slightly conservative* ($value = 3$), their cosine similarity ($sim = 0.98$) would be higher than a pair of students with the same political orientations but where one student is *Not at all interested* ($value = 0$) and the other is *Very interested* ($value = 3$, $sim = 0.7$). Each category has a different number and range of attributes, and the similarity scores vary accordingly, however the magnitude is consistent and allows for graph correlation analysis, as described next.

To find the relationship between the multiplexity (mw_{ij}) and the profile similarity of two individuals, a standard matrix correlation coefficient is used. Given two generic graphs represented by the NxN weighted adjacency matrices A^a and A^b, the correlation coefficient per node C_i can be defined as follows:

$$C_i = \frac{\sum_{j=1}^{N} w_{i,j}^a w_{i,j}^b}{\sqrt{\sum_{j=1}^{N} w_{i,j}^a \sum_{j=1}^{N} w_{i,j}^b}} \tag{4.5}$$

where $w_{i,j}^a$ represents the multiplex weight between a pair of nodes in layer a. From the definition of the correlation coefficient per node, the graph correlation coefficient which measures the correlation between the two weighted matrices can be derived as follows:

$$C(A^a, A^b) = \frac{\sum_{i}^{N} C_i}{N} \tag{4.6}$$

political	music	health	situational
0.6**	0.49*	0.6**	0.56*

Table 4.3: Graph correlations between multiplex weight and each similarity score, p-value < *0.001* **, *0.01* *.

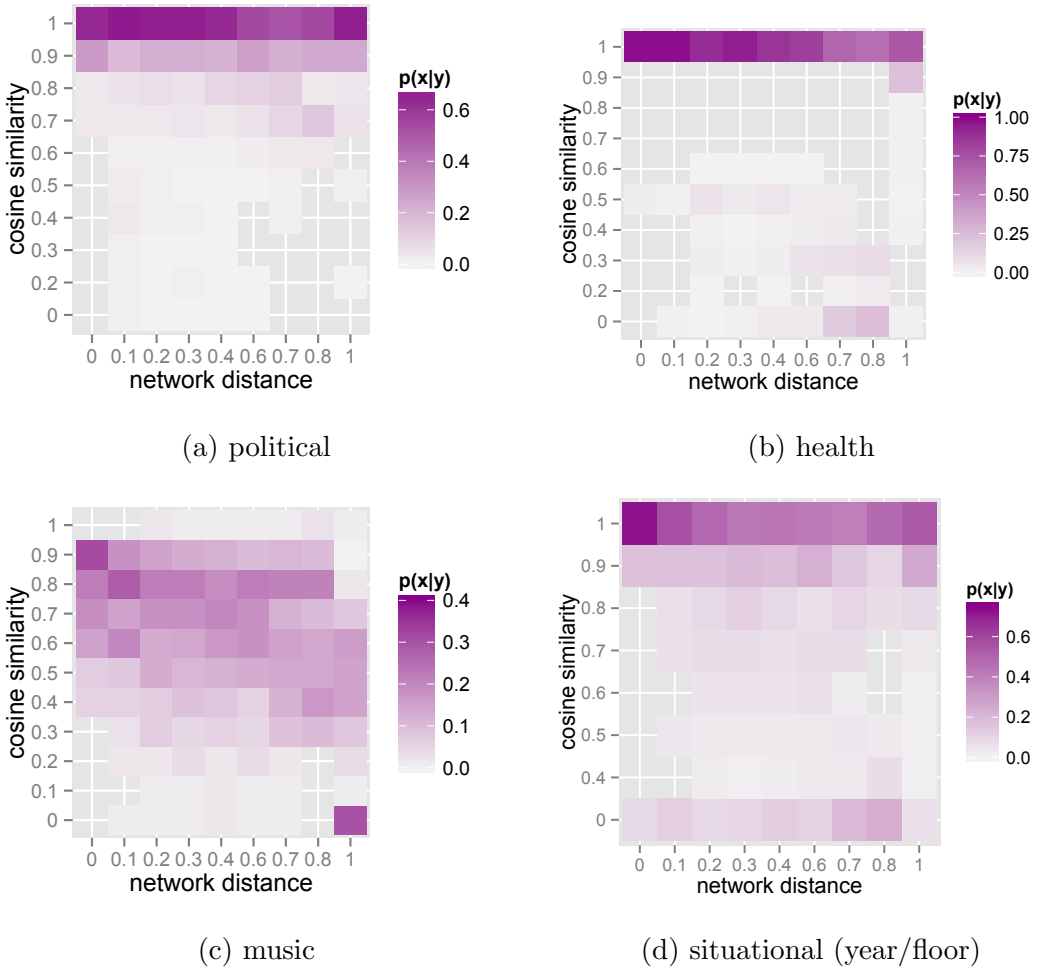
which is essentially the average correlation coefficient for all nodes. The graph correlations are calculated between the adjacency matrix of the multiplex and that of the pairwise similarity per category. In essence, each mw_{ij} is compared with its correspondent similarity weight, and then the average for the whole graph is taken (see Table 4.3). We find that there is a significant positive relationship between the multiplex edge weights and the similarity across categories.

The highest correlations are with the political and health factors (C = 0.6 for both), signifying that these are most closely related to the multiplex tie strength, followed by situational factors (C = 0.56), and music (C = 0.49). This means that multiplex ties tend to be observed in conjunction with high profile similarity.

Next, the Spearman rank degree correlations are computed between the weighted multiplex degree and weighted similarity degree of each node and the effects on a neighbourhood level are observed in Fig. 4.4. Those nodes with a high similarity degree also have a high multiplexity degree in consistence with the graph correlations on a per edge basis. This signifies that students who have more multiplex ties are also more similar to their neighbours than less popular ones in terms of degree rank. From the correlations on a per edge level, along with correlations on the neighbourhood level, it can be confirmed that *the greater the multiplexity, the greater the similarity observed across categories.*

Homophily & Multiplexity Homophily is a network phenomenon, which is distinct from homogeneity in that it implies the occurrence of non-random similarity in pairs of connected nodes whereas homogeneity is simply similarity between pairs. The presence or absence of homophily in the student community can be measured as a function of network distance and profile similarity. Here, distance is defined as the standard weighted network distance. The network is weighted using the multiplex weight mw_{ij}. The distance is then equivalent to the shortest path between two nodes in the network. A distance of 0 is indicative of full connectivity in the multiplex. This means that the pair is connected on all three layers. Direct connectivity of one hop exists up to a multiplex weight of 0.2, where weights are normalised in the range 0-1. A distance of 1 represents a non-existent path between two nodes.

With the expectation that individuals at a shorter distance in the multiplex network are more similar than those further away, the conditional probability of the similarity between pairs of connected students given a specific network distance, is measured as shown in Fig. 4.5. At first glance, we can distinguish between the top graphs (4.5a and 4.5b)



(a) political

(b) health

(c) music

(d) situational (year/floor)

Figure 4.5: Conditional probability of similarity given a certain network distance ($P(x|y)$) in terms of multiplex weighted shortest path. We consider the entire multiplex networks, where the maximum distance was 0.9, and 1 is where no path exists between two nodes.

as having consistently high probability of high similarity over distance, and the bottom graphs (4.5c and 4.5d) as having a diagonal distribution of high probability, with high similarity probability decreasing over distance.

The first two figures show an overall homogeneity in terms of political and health factors. At a distance of 1 (non-connected nodes), the probability of high similarity is still high, indicating that two nodes with high multiplexity and high similarity in these categories could be connected at random. High homogeneity can be expected in the study, given that students are co-residing and share the same context.

On the other hand, homophily exists where there is non-random similarity between individuals with shorter multiplex distance. This is most evident in subfigure 4.5c - music preferences, where there is a clear shift in high probability from top left to low right as distance increases. This means that those pairs at a short distance in the network, are also highly likely to have a similar taste in music (sim = 0.9, where $dist$ = 0), whereas those pairs who are further from each other in the network have a lower similarity in

music taste. For unconnected pairs, the similarity is especially low (near 0), indicating that edges in the network with respect to music are non-random and highly dependent on musical preferences.

Fig. 4.5d on the other hand shows an interesting divide between low and high similarities. Most pairs are grouped into very high or very low similarity, and appear at the top row and bottom row of the graph. Therefore, students tend to be either in the same year and floor or in different years and different floors, which may be as a result of room allocation according to year. There is a high chance ($P = 0.7$) that those who live and study together ($sim = 1$), also have a highly multiplex tie, represented by the distinguishable top left tile at position (0,1), and less so for those who live further apart and/or are in a different year.

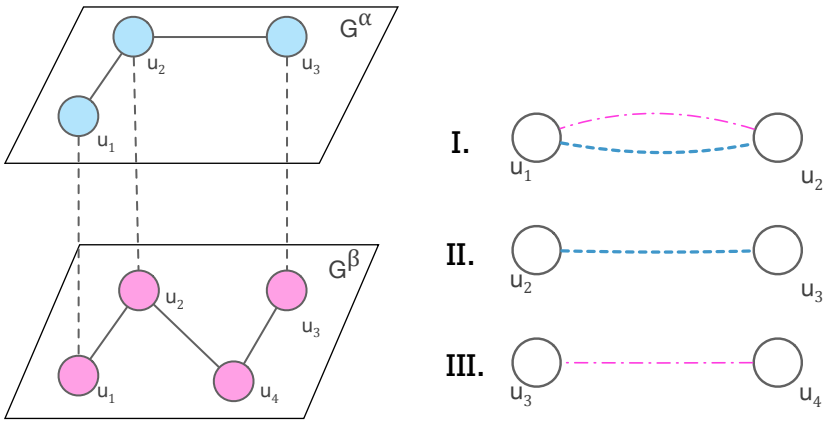
Despite the community being highly homogeneous in political and health aspects, *the shorter the distance in the multiplex network, the greater the similarity between two nodes* in the music and situational categories, which is indicative of homophily. In conclusion, a multilayer approach to social network tie strength and homophily can be beneficial to understanding the social dynamics of a student community. In the following section, we will show that the concept of multiplexity can also be useful in online social network link prediction, demonstrating that multilayer frameworks can be a useful tool for social bootstrapping and friend recommendations due their comprehensive perspective on the online social "ecosystem".

4.2 Link Prediction in Geo-Social Multiplex Networks

Link prediction systems are key components of social networking services due to their practical applicability to friend recommendations and social network bootstrapping, as well as to understanding the link generation process. Link prediction is a well-studied problem, explored in the context of both OSNs and location-based social networks (LBSNs) [106, 125, 47, 165]. However, only very few link prediction works tackle multiple networks [103, 178, 156, 199], while *most link prediction systems only employ features internal to the network under prediction*, without considering additional link information from other OSNs. In this chapter, we will examine the problem of link prediction in light of a multilayer approach and set the prediction task across the geographical and social layers of two distinct and heterogeneous platforms.

4.2.1 A Multilayer Approach to Online Ties

In the online social "ecosystem", while each platform can be explored separately as a network in its own right, this does not capture the dimensionality of online social life, which spans across many different platforms. Fig. 4.6a illustrates the concept by showing



(a) Multilayer network (b) Link types

Figure 4.6: Multilayer model of OSNs with different link types: I. Multiplex link; II. Single-layer link on G^α; and III. Single-layer link on G^β.

how two graphs G^α and G^β are coupled by common members, while some links may be present or absent across the two graphs. As this represents the general case of online social networks, members need not be present at all layers and the multilayer network is not limited to two layers.

Fig. 4.6b illustrates three link types for the case of a two layer network. Firstly, a *multiplex link* between two nodes i and j can be defined as a link that exists between them *at least in two layers* $\alpha, \beta \in \mathcal{M}$. Second, a *single-layer link* between two nodes i and j exists if the link appears *only in one layer* in the multilayer social network. In systems with more layers, multiplexity can take on a value depending on how many layers the link is present on as demonstrated in the previous section of this chapter and Chapter 3. Since the multiplex model is applied to online social media here, the number of layers can be expected to remain in the single digits due to cognitive limits in human interaction [86]. This will ensure that with each additional layer, the value of link between two individuals increases and information is added to their tie strength [84].

At the beginning of the chapter, we introduced two aggregations of the multiplex graph in Equation 4.3 for the core neighbourhood and in Equation 3.2 in Chapter 3 for the union or global neighbourhood. This simple formulation allows for powerful extensions of existing metrics of neighbourhood similarity, widely used for link prediction in online social networks. The Jaccard similarity of two users i and j's global neighbourhoods can now be defined as:

$$jacc_{glob}(i,j) = \frac{|N_{\mathcal{M}}(i) \cap N_{\mathcal{M}}(j)|}{|N_{\mathcal{M}}(i) \cup N_{\mathcal{M}}(j)|} \tag{4.7}$$

where the number of *common* friends is divided by the number of *total* friends of i and j. The same can be done for the core neighbourhood $N_{\mathcal{C}}$ of two users.

The Adamic/Adar index for link likelihood [2], which takes into account the overlap of two neighbourhoods based on the popularity of common friends (originally through web pages) has been a popular link prediction feature in a single-layer network. It can be extended to the multilayer context through our definition of neighbourhoods as:

$$aa\_sim_{glob} = \sum_{z \epsilon N_{\mathcal{M}}(i) \bigcap N_{\mathcal{M}}(j)} \frac{1}{log(|N_{\mathcal{M}}(z)|)} \tag{4.8}$$

where it is applied to the global common neighbours between two nodes but can be equally applied to their core neighbourhoods. Both the Jaccard similarity and the Adamic/Adar index have been shown to be effective in solving the link prediction problem in both social and location-based networks [106, 165]. We have extended these and apply them to the multilayer domain to predict online social links across and between Twitter and Foursquare – two heterogenous social networking platforms.

4.2.2 Data

Twitter and Foursquare are two of the most popular social networks, in terms of user base size and interest in the research community. They have distinct broadcasting functionalities - microblogging and venue check-ins. While Twitter can reveal a lot about user interests and interactions, Foursquare check-ins provide a proxy for human mobility. In Foursquare users check-in to *venues* that they visit through their location enabled devices, and share their visits of a place with their connections. Foursquare is two years younger than Twitter and its broadcasting functionality is exclusively for mobile users (50M to date[1]), while also 80% of Twitter's 284M users are active on mobile[2]. Twitter and Forusquare generally allow anyone to "follow" and be "followed", where followers and followed do not necessarily know one another. An undirected relationship can be constructed from this, where a link can be considered between two users if they both follow each other reciprocally [99].

The dataset was downloaded from the public Twitter and Foursquare APIs between May and September 2012 for three major US cities, where tweets and check-ins were downloaded for users who had checked in during that time, and where those check-ins were shared on Twitter. We initially identified Foursquare users on Twitter by hashtags that pertain to the Foursquare service and then continuously downloaded their tweets over the four month period. Therefore, our dataset contains a subset of Foursquare users who publicly share their check-ins via the Twitter service, who are estimated to be 20-25% of the Foursquare user base [137]. This allows for the study of the intersection of the two networks through users who have accounts and are active on both Twitter and

[1] `https://foursquare.com/about`

[2] `https://about.twitter.com/company`

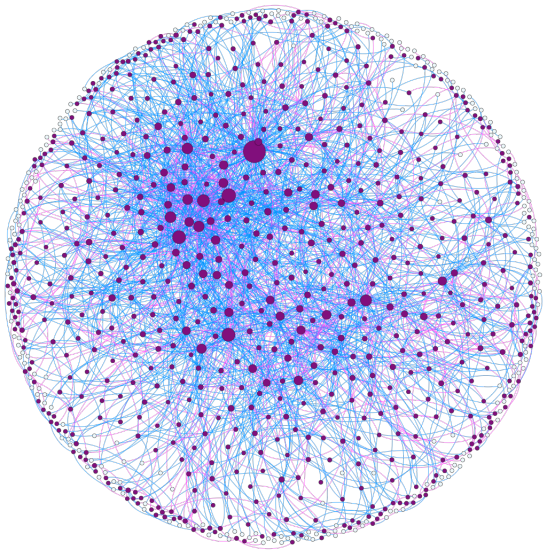


Figure 4.7: Social network graph for San Francisco. Blue edges are single-layer edges, while pink edges are multiplex edges. The node size is proportional to the degree in the union aggregation of that node.

Property	New York	Chicago	SF	All
$\|V^{\mathcal{M}}\|$	6,401	2,883	1,705	10,989
$\|E^{T\cap F}\|$	9,101	5,486	1,517	16,104
$\|E^{T\setminus F}\|$	13,623	7,949	1,776	23,348
$\|E^{F\setminus T}\|$	6,394	4,202	863	11,459
$< k_{glob} >$	4.55	6.12	2.44	4.63
$< k_{core} >$	1.42	1.9	0.89	1.47
tweets	2,509,802	1,288,865	632,780	4,431,447
checkins	228,422	105,250	46,823	380,495
venues	24,110	11,773	6,934	42,817

Table 4.4: Dataset properties: number of users (nodes); number of multiplex links (edges); number of Twitter and Foursquare only edges; average global and core degrees; activity and venues per city.

Foursquare. Tweets were divided into *check-ins* and *tweets* depending on whether the content of the tweet was a Foursquare check-in or not. A tweet is in the form (*userId, mentions, hashtags*), where the actual content of the tweet is not considered apart from whether it mentions another user or identifies with a topic through Twitter's hashtag (#) paradigm of topics. Check-ins are in the form (*userId, venueId, coordinates, timestamp*) where the temporal and spatial aspects of the check-in are taken into account and not its semantic properties. At the end of the period, the social network of each user was also downloaded from both platforms by obtaining the user ids of their followers and who they are following as well as Foursquare friends of up to one hop in the network. We believe the dataset does not contain bots or other automated accounts as only real users are known to post content through Foursquare due to its mobile application context.

Table 4.4 shows the details for each city, in terms of activity and venues, multilayer

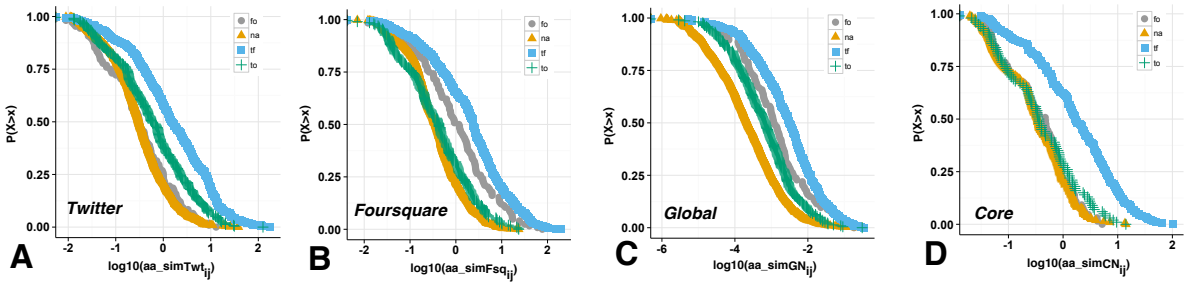
edges and degrees for each network, where $E^{T \cap F}$ denotes the set of edges, which exist on both Twitter and Foursquare, $E^{T \setminus F}$ and $E^{F \setminus T}$ are the sets of edges on Twitter only and Foursquare only, respectively. Fig. 4.7 additionally illustrates the case of San Francisco, where blue edges represent single-layer links on either Foursquare or Twitter, and pink edges represent multiplex links on both. A Fruchterman Reingold graph layout [73] is used to show the core-periphery structure of the network, with larger nodes having a larger degree in the union aggregation.

4.2.3 Properties of Online Multiplex Links

The first goal at hand is to gain insight into the geo-social structural and interaction properties of multiplex links in the multilayer online social network and how they differ from other link types. We study the three types of links as described in the multilayer model above: multiplex links across both Twitter and Foursquare, which are denoted as *tf* for simplicity; single-layer links on Foursquare only (denoted as *fo*); single-layer links on Twitter only (denoted as *to*). These are then compared to unconnected pairs of users (denoted as *na*). The insight gained from the discriminative power of each feature, can be used to interpret the results of the link prediction tasks defined in the following section.

Link Multiplexity and Structural Similarity The number of common friends between two individuals has been shown to be an important indicator of a link in social networks [106]. Moreover, the neighbourhood overlap weighted on the popularity of common links between two users has been shown to be a good predictor of friendship in online networks [2]. Fig. 4.8 shows the cumulative distribution of the Adamic/Adar index of neighbourhood similarity across the various single and multilayer configurations of the networks at hand and each of the four link types. Figs. 4.8a and 4.8b show the cumulative distribution over the single-layer configurations of Twitter and Foursquare respectively, while Figs. 4.8c and 4.8d show the distribution over the core and global multilayer configurations. These plots allow us to reason about the fraction of pairs of users with an Adamic/Adar index greater than a certain threshold which relates to the way that features are ranked in a machine learning framework.

Each figure shows the fraction of Adamic/Adar indices greater than the given threshold. In Fig. 4.8a we can see that 25% of Twitter user pairs (to) have an overlap of $10^{0.3}$ or greater, while 25% of multiplex tie pairs (tf) have 10^1 or higher. Those pairs that are not connected (na) and those which are only connected on Foursquare (fo) have a similarly lower Adamic/Adar threshold of 10^0. The results over different fractions of user pairs remain consistent where multiplex tie pairs (tf) always have a higher Adamic/Adar index threshold than Twitter only (to), Foursquare only (fo) and no link (na) pairs, based on the CCDF curves. These results are analogous for the Foursquare network, where we can observe an Adamic/Adar index of approximately 10^1 for 25% of multiplex pairs, closely



(a) Twitter (b) Foursquare (c) Global (d) Core

Figure 4.8: Complementary cumulative distribution function of the log Adamic/Adar index for the different network configurations, grouped by link type - Twitter overlap (A), Foursquare overlap (B), Global overlap (C), Core overlap (D). Each figure shows the fraction of links with an aa_sim value greater than x.

followed by Foursquare only (fo) pairs and then Twitter only (to) and na user pairs with a value of 10^0. From the two single layer configurations, we can see that multiplex links exhibit higher structural similarity at each threshold, followed by links native to the platform and then those exogenous to the platform and finally unconnected user pairs.

With respect to the multilayer configurations, it can be observed in Fig. 4.8C that 50% of user pairs which are not connected have an Adamic/Adar index of 10^{-4} or greater, whereas 50% of single-layer links (fo and to) have 10^{-3} or higher and finally multiplex link pairs (tf) have an index of $10^{-3.5}$ or greater. On the other hand, in the core configuration in Fig. 4.8d we can see a division between multiplex link types and all other link types, where 25% of all pairs of all multiplex ties (tf) have an index of approximately 10^1 or greater while all other link types have a lower threshold of 10^0 or higher. While this is somewhat expected, it shows that the core configuration is a good proxy for detecting multiplex ties. In agreement with previous studies of tie strength [76], multiplex links share greater structural similarity than other link types across network configurations and this will be a useful property in our link prediction problem.

Link Multiplexity and Interaction The volume of interactions between users is often used as a measure of tie strength [139]. In this section, the volume of geo-social interactions on Twitter and Foursquare are shown to discriminate between the presence of the various link types. A number of interaction features are extracted from the two services, which along with the previously introduced structural features, will be examined in the following section in light of their predictive power. These interaction features are:

- *Number of mentions:* The number of instances in the dataset in which user i has mentioned user j on Twitter during the period. Mentions include direct tweets and retweets mentioning another user. Any user on Twitter can mention any other user and does not have to be following that user in the social network. This allows for

this feature to be measured across pairs which do not have a link on any network (na). Twitter users have been shown to exhibit favouritism for a small group of their contacts when it comes to mentions (retweets) [99].

- *Number of common hashtags:* Similarity between users on Twitter can be captured through common interests. Topics are commonly expressed on Twitter with hashtags using the # symbol. We therefore measure the number of instances in which user i and user j have posted a tweet using the same hashtag. Similar individuals have been shown to have a greater likelihood of having a tie through the principles of homophily [123].

- *Number of colocations:* The number of times two users have checked into the same venue within a given time window. In order to reduce false positives, a shorter time window of 1 hour only is considered. Two users who appear at the same place, at the same time on multiple occasions, have a higher likelihood of knowing each other (and therefore having a link on social media). Each colocation is weighted on the popularity of a place in terms of the total user visits, to reduce the probability that colocation is by chance at a large hub venue such as an airport or train station. The importance of colocations has been highlighted in discovering social ties as well as place-focused communities [30].

- *Distance:* Human mobility and distance play an important role in the formation of links, both online and offline, and have been shown to be highly indicative of social ties and informative for link prediction [193]. The distance between the geographic coordinates of two users' most frequent check-in locations were calculated as the Haversine distance, the most common measure of great-circle spherical distance: $dist_{ij} = haversine(lat_i, lon_i, lat_j, lon_j)$, where the coordinate pairs for i, j are those of the places where users with more than two check-ins have checked in most frequently, equivalent to the mode in the multiset of the venues where they have checked in. This allows for minimising data loss while increasing the probability that a most frequent location will emerge, similar to previous related work in the field [44, 163, 136].

Two additional features are considered, which merge information from the Twitter social network and the Foursquare location network. In order to capture the tie strength between a pair of users in the multilayer network, their similarity based on the social layer can be considered, or the *number of common hashtags*, denoted by sim_{ij} and their spatial similarity, or the *distance* between their most frequented venues on Foursquare, denoted by $dist_{ij}$. Inspiration for this measure is drawn from gravity models in transportation studies where the attraction between two entities is proportional to the importance of their interaction over their distance [155]. The global similarity can be defined as the Twitter similarity over Foursquare distance as:

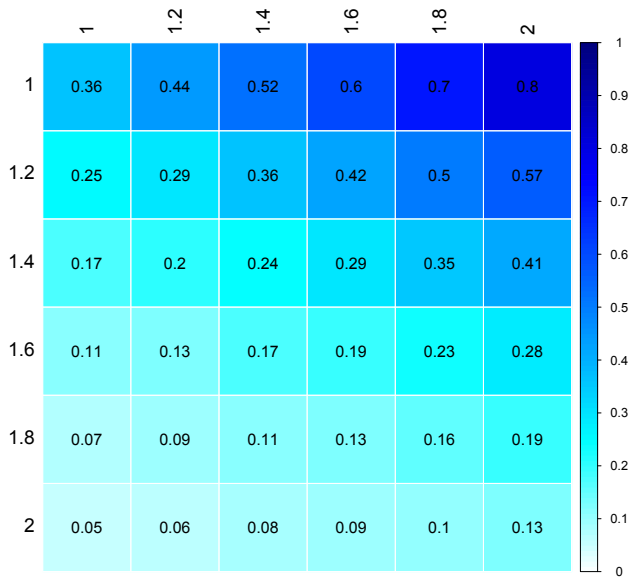


Figure 4.9: Exponent matrix for sim_{glob}. Colour gradient indicates the optimal exponents in terms of difference maximisation between the medians of the multiplex and non-existent link types - $|Md_{tf} - Md_{na}|$.

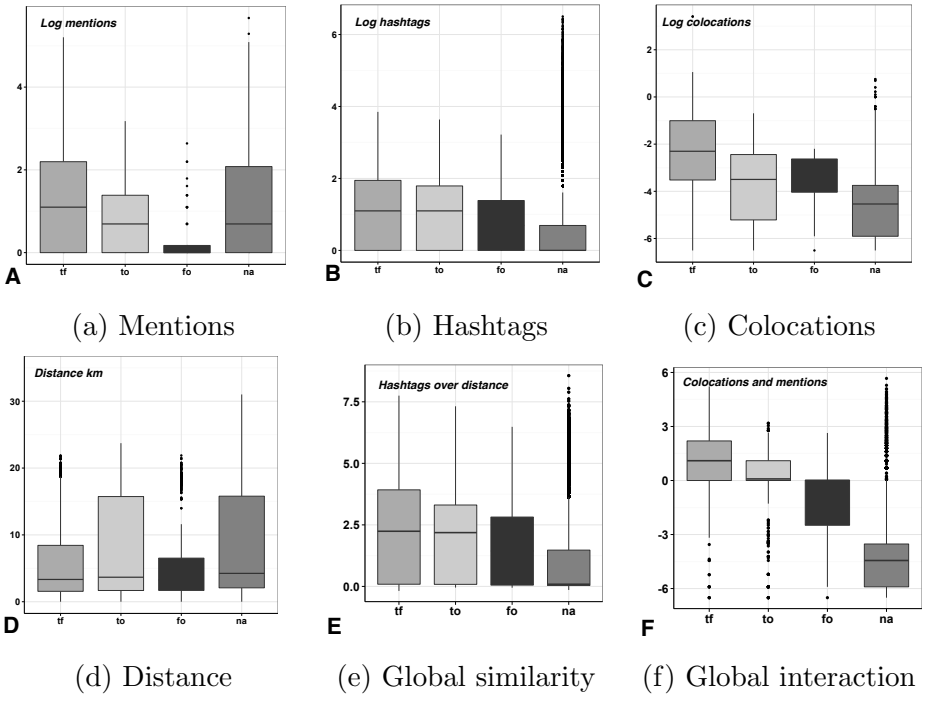
$$sim_{glob}(i,j) = \frac{sim_{ij}^{a}}{dist_{ij}^{b}} \tag{4.9}$$

where exponents a, b are chosen based on the context of at hand. In this case, a is the potential for the similarity measure to reflect a reciprocal link between two users, whereas b is a parameter related to how well connected the two venues are and therefore how significant the distance between them is, similar to the gravity model's original use in transportation [155]. The exponents $a = 2, b = 1$ are set after optimising for the exponents that maximise the difference between the median values of multiplex links (tf) and no link (na). Figure 4.9 shows how these results vary across different exponents a and b in the range [1,2].

The second feature which captures the complete interaction across layers of social networks can be defined as:

$$int_{glob}(i,j) = \sum_{\alpha}^{M} k^{\alpha} |int_{ij}^{\alpha}| \tag{4.10}$$

where int can be any type of interaction between i and j in layer α and interactions are summed across layers and weighted by a constant k for each layer. This allows for adjustments based on the weighted importance of an interaction, specific to the context of the layer. In our case we consider mentions and colocations as the interactions across



(a) Mentions (b) Hashtags (c) Colocations

(d) Distance (e) Global similarity (f) Global interaction

Figure 4.10: Interaction features' distribution for each link type. Panel Figure A-C show the distributions of Twitter mentions, Common hashtags, and Number of colocations in log scale, distribution of distance in km between the home locations of users according to the type of link they have (top 10% of distances are excluded for figure readability), and the distribution of the multilayer similarity and interaction features.

layers and a coefficient $k = 1$ for both layers as we would like to maintain the empirical properties of interactions and after optimising for a number of different coefficients.

In Fig. 4.10, the four types of spatial and social interaction on the two social networking services as well as the two multilayer geo-social features are presented in the order in which they were presented. Each box-and-whiskers plot represents an interaction between multiplex links (tf), Twitter only (to), Foursquare only (fo), and unconnected pairs (na) on the x axis. On the y axis we can observe the distribution divided in four quartiles, representing 25% of values each. The dark line in the middle of the box represents the median of the distribution, while the dots are the outliers, where the definition for an outlier is a value which is less than the first quartile or greater than the third quartile by more than 1.5 times the interquartile range between quartile 3 and 1. The "whiskers" represent the top and bottom quartiles, while the boxes are the middle quartiles of the distribution.

In terms of Twitter mentions (Fig. 4.10a), multiplex ties (tf) exhibit higher values of mentions than any other group, including the Twitter only group (to) with a median value of 10^1 and top-quartile values above 10^4. Pairs of users connected only on Foursquare (fo) do not typically mention each other on Twitter although this is made possible by the service. On the other hand, mentions are just as common between users who are not connected on any network (na) as between those who are connected on both (tf), which may be as a

result of mentioning celebrities and other commercial accounts. This, however, is not the case for *hashtags* (Fig. 4.10b), where we find that almost all of unconnected users share 10 hashtags or less with the exception of outliers. While *mentions* are more discriminative between multiplex links (tf) and single-layer connectivity (to and fo), *hashtags* are better at distinguishing between links and non-links (na) in terms of median values.

With regard to Foursquare spatial interaction in Figs. 4.10c and 4.10d, multiplex ties (tf) have the highest probability of multiple *colocations* with a median value of $10^{-3.8}$. Despite being weighted by the popularity of a venue, values in the top quartile of unconnected pairs (na) are relatively high with respect to other link types. However, in terms of median values there is still a distinction between the different levels of multiplexity which each link type represents. On the other hand, while *distance* (Fig. 4.10d) does not vary much in terms of median values for the different link types, based on the top quartiles of the distributions across link types, it appears that Foursquare only pairs (fo) are more likely to frequent locations close to each other, closely followed by multiplex link pairs (tf) where distances for both are below 20km. Twitter only (to) and unconnected pairs frequent locations similarly further away. This indicates that both Foursquare spatial features are better at distinguishing multiplex links and native Foursquare links than other link types based on the distributions observed in agreement with previous literature which has suggested that geographical features are powerful social link predictors [30].

In Figs. 4.10e and 4.10f we can compare the multilayer geo-social features we defined above to the single-layer social and geographic features observed. The distribution of the sim_{glob} measure, integrating similarity and distance as factors of attraction between pairs of users, can discriminate between link types mainly based on the maximum value in the top quartile of the distributions in Fig. 4.10e, where we observe that the maximum values for multiplex links are higher than any other link type (over 7.5), whereas the maximum value for unconnected pairs is approximately 4 while the median is 0. This shows that only values with low similarity and high distance fall below 0, whereas most pairs of users have less negligible similarity where values around 1 indicate a balance between distance and similarity.

In Fig. 4.10f the distinction between different link types in the distributions of values is more striking than for any of the single-layer features. We can see that each median value is significantly different – multiplex links (tf) are the highest with a median of $10^{1.5}$, followed by to links (10^{0}), fo links ($10^{-1.5}$) and finally non-present links (10^{-4}). This satisfies two desirable properties for link prediction – distinct thresholds between link types, and a discriminative threshold between the non-existent links (na) and all other link types, on which to base binary decisions of the presence/absence of a link.

4.2.4 A Multilayer Approach to Link Prediction

Having empirically shown the value of the different features in distinguishing between different link types above, here the question of how this information can be used to predict links across layers of social networks is explored. The likelihood of forming a social tie as a process that depends on a union of factors, using the Foursquare, Twitter, and the multilayer features defined up until now is evaluated in a supervised learning approach, and their predictive power is compared in each feature set respectively.

Prediction Space The main motivation for considering multiple social networks in a multilayer construct is that each layer carries with it additional heterogeneous information about the links between the same users, which can potentially enhance the predictive model. In the present context there are two distinct layers of information - the spatial movements of users from Foursquare and their parallel social interactions on Twitter. This evaluation explores whether by using spatial features from one network layer (Foursquare), it is possible to predict links on the social network layer (Twitter), and vice versa. In light of the multilayer nature of OSNs, it is also explored whether better prediction can be achieved by combining features from multiple networks.

Formally, for two users in the multilayer network $i, j \in \mathcal{M}$, where $V^{\mathcal{M}}$ are the nodes (users) that are present in any layer of the multilayer network, we employ a set of features in a supervised learning framework that output a score r_{ij}^{α} so that all possible pairs of users $V^{\mathcal{M}} \times V^{\mathcal{M}}$ are ranked according to their expectation of having a link e_{ij}^{α} on a specific layer α in the network. Two distinct prediction tasks can be specified:

(1) Rank pairs of users based on their interaction on one network layer in order to predict a link on the other. This entails (a) training on spatial mobility interactions to predict social links on Twitter, and (b) training on social interaction features on Twitter to test on Foursquare links.

(2) Rank pairs of users based on their interaction on both network layers in order to predict a link across both (a multiplex link). We train on three sets of features – spatial interactions, social interactions, and multilayer features which are summarised in Table 4.5.

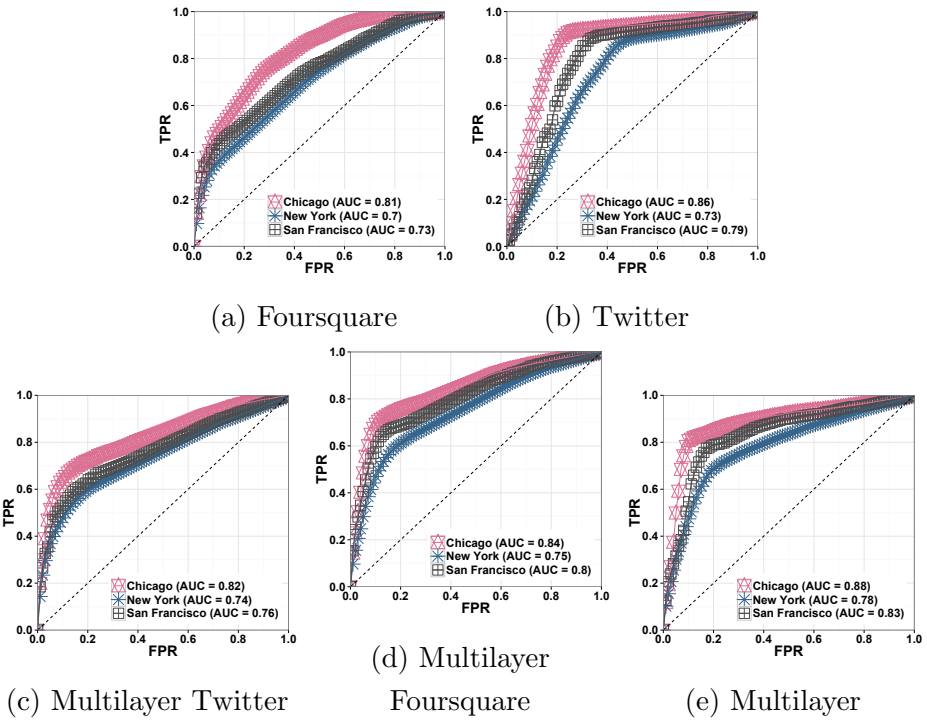
The evaluation is performed on the three datasets described in Table 4.4 for the cities of San Francisco, Chicago, and New York to show performance on these tasks across urban geographies. In terms of algorithmic implementation, public versions of the algorithms available in [146] were used. Supervised learning methodologies have been proposed as a better alternative to unsupervised models for link prediction [108]. The data is fit to a Random Forest classifier [25], which uses a sub-sampling and averaging technique across a number of tree estimators to improve the predictive accuracy and control over-fitting. Subsampling takes place with replacement and is equal to the training set size. Each

Twitter features	
$mentions$	$\lvert mentions_{ij} \rvert$
$hashtags$	$\lvert hashtags_{ij} \rvert$
$jacc$	$\frac{\lvert N_i^T \cap N_j^T \rvert}{\lvert N_i^T \cup N_j^T \rvert}$
aa_sim	$\sum_{z \in N_i^T \cap N_j^T} \frac{1}{log(\lvert N_z^T \rvert)}$
Foursquare features	
$colocs$	$\lvert colocations_{ij} \rvert$
$dist$	$haversine(lat_i, lon_i, lat_j, lon_j)$
$jacc$	$\frac{\lvert N_i^F \cap N_j^F \rvert}{\lvert N_i^F \cup N_j^F \rvert}$
aa_sim	$\sum_{z \in N_i^F \cap N_j^F} \frac{1}{log(\lvert N_z^F \rvert)}$
Multilayer features	
int_{glob}	$\sum_{\alpha}^{M} \lvert int_{ij}^{\alpha} \rvert$
sim_{glob}	$\frac{sim_{ij}^a}{dist_{ij}^b}$
$jacc$	$\frac{\lvert N_C(i) \cap N_C(j) \rvert}{\lvert N_C(i) \cup N_C(j) \rvert}$
aa_sim	$\sum_{z \in N_C(i) \cap N_C(j)} \frac{1}{log(\lvert N_C(z) \rvert)}$

Table 4.5: Summary of link features. We denote the Twitter neighbourhood as N^T and the Foursquare neighbourhood as N^F.

prediction task is optimised across two parameters: the number of tree estimators and the max depth allowed for each estimator.

A 10-fold stratified cross-validation testing strategy is used in addition: for each test we train on 90% of the data and test on the remaining 10% and each fold set contains approximately the same percentage of samples of each target class as the complete set since the number of prediction items in the data are in the order of $|V^{\mathcal{M}}|^2$. For every test case, the user pairs are ranked according to the scores returned by the classifier for the positive class label (i.e., for an existing link), and subsequently, all possible thresholds of probability in terms of true positive (TP) and false positive (FP) values rate are plotted against each other as Receiver Operating Characteristic (ROC) curves. The Area Under the Curve (AUC) scores are used from these curves to report the relative performance of each task by averaging the results across all folds, where the fraction of positive examples correctly classified is taken into account as opposed to the fraction of negative examples incorrectly classified. ROC analysis can provide insight about how well the classifier can be expected to perform in general, at a variety of different class imbalance ratios and therefore, against different random baselines that could correspond to these ratios.



(a) Foursquare (b) Twitter

(c) Multilayer Twitter (d) Multilayer Foursquare (e) Multilayer

Figure 4.11: ROC curves for the Random Forest classifier and Area Under the Curve (AUC) scores for each city dataset.

Multilayer Link Prediction The evaluation is presented using ROC curves and the corresponding Area Under the Curve (AUC) scores across cities, shown in Fig. 4.11. First, training is performed on the Twitter social interaction features summarised in Table 4.5 and tested on the Foursquare target labels. Formally, for a pair of users i and j a feature vector $\mathbf{x}_{\mathbf{ij}}^{\alpha}$ encoding the values of the users' feature scores on layer α in the multilayer network can be defined. A target label $y_{ij}^{\beta} \in \{-1, +1\}$ is also specified, representing whether the user pair is connected on the β layer under prediction. The supervised Random Forest classifier (best performance achieved with 45 tree estimators, allowing for a maximum tree depth = 25 each) is used to predict links from one layer using features from the other.

Fig. 4.11a shows the ROC curves and respective AUC scores for each city in predicting Foursquare links from Twitter features, ranging between 0.7 for the New York dataset to 0.81 for Chicago, and 0.73 for San Francisco. These results represent the probability that the classifier will rank higher a randomly chosen positive instance than a randomly chosen negative instance [69]. On the other hand, we can consider the reverse task of predicting Twitter links using Foursquare features in Fig. 4.11b, where an AUC scores of 0.86, 0.73, and 0.79 are obtained for the three cities respectively. Slightly higher results can be observed for Twitter links, which may be a result of the higher number of Twitter links in the dataset or as a result of the greater difficulty of the inverse task. These results are compared to the traditional single-layer prediction task of Twitter links from Twitter features and Foursquare links from Foursquare features internal to the platform where an AUC= 0.86 and AUC= 0.88 are achieved on average between cities with the same

Random Forest set-up. This shows that our performance across services is comparable to that within the service itself.

It has been observed in the preceding analysis on link types that those pairs connected only on Foursquare do not exhibit strong interaction on Twitter by exchanging a low number of mentions and having low neighbourhood overlap, however, those pairs of users connected on both platforms, exhibit high interaction across. It can therefore be expected that a large number of stronger multiplex ties have been identified in this task. In the second prediction task, this assumption is tested by observing if it is possible to achieve higher predictive power across cities when testing on the presence or absence of a multiplex link. Formally, given a feature vector $\mathbf{x_{ij}}$, the goal would be to predict a target label $y_{ij} \in \{-1, +1\}$, where a link exists on both layers (+1) or not (-1). In Fig. 4.11c and 4.11d we can observe that it is possible to achieve greater predictive power using Twitter features in predicting multiplex links than Foursquare links in Fig. 4.11a and in using Foursquare features in Fig. 4.11b, with the highest AUC scores of 0.82 and 0.84 for each set respectively. It should be noted that the Foursquare spatial features perform slightly better than the social interaction features for Twitter, which places importance on the discriminative power of spatial interactions as also observed in the first part of the analysis. This confirms the assumption that multiplex links are easier to identify than single layer links by using the same algorithmic set-up and shows that the strength of multiplex ties exhibited in the first part of our analysis can be used to predict links across networks.

Finally, it can be observed that using multilayer and geo-social features which employ both spatial and social interactions from the two heterogeneous platforms can outperform both single layer sets in predicting multiplex links (highest AUC = 0.88 for Chicago). It is intuitive that when using information from both layers the prediction of multiplex links becomes easier and it is often the case that such multilayer network data is not available. However, it was observed that relatively good results can also be achieved using only social or only geographic information.

In order to evaluate the information added by the proposed features as compared to the previously widely used Adamic-Adar and overlap metrics, the prediction results thus far are compared with a simplified model using the Adamic-Adar and overlap features alone, while using the same predictive framework, and the change in average AUC scores between cities is computed. For the first prediction task of using the Twitter social layer features to predict links on the spatial Foursquare layer, an AUC score of 0.68 is achieved when using *aa_sim* and *overlap* features alone as compared to AUC=0.8 when using the full feature set including interactions. For the second task of using the Foursquare spatial features to predict links on the Twitter social layer, an AUC score of 0.65 was obtained when using the two structural features alone as opposed to AUC=0.75 on average across cities when using the full model. This indicates that the additional interaction features add significantly to the predictive power of the model.

When predicting the presence of a multilayer link between pairs of users, using the structural Adamic-Adar and overlap features alone, an AUC of 0.7 is achieved for the social Twitter layer, 0.71 for the spatial Foursquare layer, and 0.69 for the multilayer configuration. When compared to the full feature model (AUC=0.77, 0.8, and 0.83 respectively), a significant improvement can be noted in terms of predictive power. In conclusion, the information added by the multilayer interaction features results in a significant improvement over the existing methods based on popular structural features alone.

4.3 Related Work

As touched upon in Chapter 3, and further elaborated in the present chapter, media multiplexity [83] is the principle that tie strength is observed to be greater when the number of media channels used to communicate between two people is greater (higher multiplexity). It is a well studied property in the social sciences [84] and it has been explored in social networks from Renaissance Florence [142] to the Internet age [83]. In [84] the authors studied the effects of media use on relationships in an academic organisation and found that those pairs of participants who utilised more types of media (including email and videoconferencing) interacted more frequently and therefore had a closer relationship, such as friendship. The strength of social ties is an important consideration in friend recommendations and link prediction [76] but has been previously understudied in the context of multiplexity properties.

Social network research has further focused on the effects of homophily expressed in interactions *online*, where findings suggest that most of the content shared comes from weak and diverse ties. Recent research on homophily [*lit.* love of same], used sensors for tracking mobility and interactions *offline*, and showed that physical exercise, residential sector, and on-campus activities are the most important factors for the formation of social relationships, placing emphasis on spatio-temporal activities [59]. Furthermore, dynamic homophily based on political opinion was studied in the same context during the 2008 US presidential election by means of Bluetooth scanning: the researchers observed increased proximity around the presidential debates between students with the same political orientation [118]. The authors in [10] showed that in the context of Facebook, while strong ties are consistently more influential, weak ties are collectively more important and users consume and share information produced largely by those with whom they interact infrequently. Diversity in online social network exchanges has also been observed in Twitter, where users re-tweet more content from topically dissimilar ties [116].

As we discuss in Chapter 1, online social media has become an ecosystem of overlapping and complementary social networking services, inherently multiplex in nature, as multiple links may exist between the same pair of users [96]. In this section, multiplex tie strength defined in the first part of this chapter is leveraged through the geographic and social

interactions of users and applied to the classic networks problem of link prediction [106]. Unlike previous work [18, 144, 156], the multilayer link prediction problem was framed across online social network platforms and media multiplexity was applied as a measure of tie strength, showing its applicability to link prediction in the geo-social domain.

The problem of link prediction in online social networks has been actively researched in the past decade, following its ignition by the seminal work of Liben-Nowell and Kleinberg [106]. Since then, it has been applied to various platforms and services. For instance, in [165] the authors exploit place features in location-based services to recommend friendships and in similar spirit the authors in [158] show how using both location and social information from the same network significantly improves link prediction, while in [8] a new model based on supervised random walks is proposed to predict new links in Facebook. Link prediction has also been approached in the multidimensional setting [156] and in multi-relational networks [199], however, these works build on features that are endogenous to the system that hosts the network of users. Drawing upon these works, the present evaluation tests on heterogeneous and fundamentally different network layers from two distinct platforms - social network Twitter and location-based social network Foursquare - by mining features from both. This approach differs in that it frames the link prediction task across layers in the context of multilayer networks, rather than partitions of the same network.

4.4 Discussion and Implications

Social media sites have had a profound effect on the way we maintain close and distant social relationships, on their number and their diversity, and the cultivation of our social capital, as previously discussed in Chapter 1 [191, 192]. Despite the vast potential for communication through social media such as Facebook, users tend to interact mostly with their closest friends [6]. Friends tend to come from similar socio-demographic backgrounds, share common interests and information. This presents evidence of homophily, or the long-standing social truth that "similarity begets friendship" [148]. While it has interesting implications in social networks in terms of link prediction [3], resilience [131], and preferential attachment [143], homophily also leads to the localisation of information and resources into socio-demographic space [123]. Conversely, diversity in social contacts has been shown to be of great importance for social and economic wellbeing, both at individual and community levels [63].

In this chapter a new multiplex measure of social tie strength was introduced and applied to uncover the presence of homophily within a community of 74 students, with respect to political orientation, music preferences, health habits, and situational factors. This methodology is inspired by Fischer's early work on homophily [70], where it was observed that the more types of relations that exist between two people (e.g., friends, kin, neigh-

bour), the stronger their bond, and the stronger the effects of homophily (similarity) between them. Although very different from kinship relations, the same principle of tie strength and depth was found in media multiplexity with regards to various types of media in the organisational environment [84], where stronger ties interact through more types of media than weaker ones.

The strength of ties manifested through multiplexity is expressed through a greater intensity of interactions and greater similarity across attributes both the offline [83], and in the online context. A number of features, which take into consideration the multilayer neighbourhood of users in OSNs were explored in this chapter. The Adamic/Adar coefficient of neighbourhood similarity in its core neighbourhood version proved to be a strong indicator of multiplex ties. Additionally, the introduced combined features, such as the global interaction and similarity over distance, discriminated more distinctively the type of link that exists between two users, than its single-layer counterparts. These features can be applied across multiple networks and can be flexible in their construction according to the context of the OSNs under consideration.

Recently, social media has been increasingly alluded to as an *ecosystem*. This allusion comes from the emergence of multiple OSNs, interacting as a system, while competing for the same resources - users and their attention. This system aspect has been addressed in this chapter by modelling multiple social networks as a multilayer online social network. Most new OSNs joining the "ecosystem" use contact list integration with external existing networks, such as copying friendships from Facebook through the open graph protocol. Copying links from pre-existing social networks to new ones results in higher social interaction between copied links than between links created natively in the platform [202]. This study proposes that augmenting this copied network with a rank of relevance of contacts using multiplexity can provide even further benefits for newly launched services.

In addition to fostering multiplexity, however, new OSNs and especially interest-driven ones such as Pinterest for example, may benefit from similarity-based friend recommendations. Here, mobility features have been applied in addition to neighbourhood similarity from Foursquare to predict links on Twitter and vice versa, highlighting the relationship between similar users across heterogeneous platforms. Similarly in [178], the authors infer types of relationships across different domains such as mobile and co-author networks. Although using a transfer knowledge framework, and not exogenous interaction features like in the present work, the authors also agree that integrating social theory in the prediction framework can greatly improve results.

Chapter 5

Social Diversity in Geo-Social Networks

This chapter builds on the previous two by extending the application of the basic multiplex model to social diversity in multilayer social networks and introduces a new model of interconnected geo-social networks which allows for cross-layer projections of social diversity between people and places. Similarly to the previous chapter, a combination of analysis techniques are used to understand brokerage and its multilayer extension to geo-social networks, initially with a light empirical validation of the phenomenon in online social networks and followed by a more in-depth network analysis approach, extending the concept to places through a human mobility study.

In the first part of the chapter, the concept of *brokerage* as a quantifiable measure of social diversity is explored, followed by its application to a multilayer online social network. The concept of structural diversity or brokerage in social networks is applied to the multilayer setting of both online and offline interactions for the first time in this chapter. The under(over)estimation of social diversity will be uncovered by applying the effective size and efficiency metrics to the multilayer setting using empirical distributions and differences between the single layer and multilayer neighbourhoods of nodes.

Furthermore, an interconnected geo-social network model is introduced and evaluated in the context of urban social diversity. Apart from brokerage, three other measures of structural and probabilistic diversity are defined and applied to urban geographies in London. The concept of the social role of a place is introduced in the interconnected network setting and evaluated against measures of social diversity in approximating deprivation indices for London. Finally, the social diversity rank of London neighbourhoods is used to identify areas undergoing processes of change such as gentrification.

5.1 Social Diversity in Multilayer Social Networks

The brokerage potential of a node in a social network provides the node with an advantageous position compared to other nodes. In particular, nodes with high brokerage potential are typically characterised as being positioned near structural holes separating otherwise disconnected pairs of nodes [36]. In this section, the notion of structural hole will be extended to a multilayer context so as to enable brokerage opportunities to arise from an individual's combined online and offline social network.

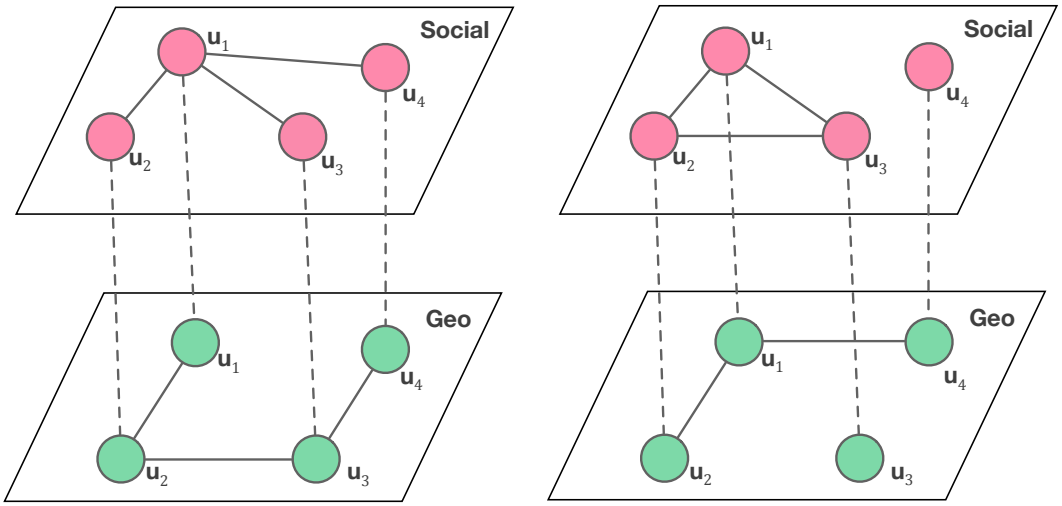
Social networks are typically represented as single layers, where nodes are connected by one type of relationship such as friendship or collaboration. While single-layer networks may be sufficient in many cases, they do not realistically capture the online and offline interaction between people, which is becoming increasingly ubiquitous. Therefore a multilayer geo-social approach to the study of social relationships is needed so as to identify the geographic layer of physical co-presence as well as the online social interaction layer. This will enable exploration of the brokerage potential of people within and across layers.

Efficiency. In Chapter 2, we introduced and formalised the brokerage of a node. An extension of this is the ego-network's efficiency, which is essentially the normalised brokerage. The efficiency E_i of node i's local neighbourhood refers to the proportion of i's neighbourhood that is non-redundant [36]. This can easily be derived by dividing the brokerage by the degree of the node:

$$E_i = \frac{S_i}{k_i}. \tag{5.1}$$

Efficiency thus helps shed light on the extent to which an individual's effort to expand social capital is directed toward novel social circles that provide exposure to non-overlapping and diverse sources of information [36].

Over(under)estimating social capital. When individuals belong to multilayer networks, such as the one in Fig. 5.1, one may over- or underestimate their brokerage potential if only a single layer is analysed (e.g., only online communication). For example, in Fig. 5.1a node u_1 brokers between nodes u_2, u_3 and u_4 in the online social network layer. There are three links connecting node u_1 to three non-redundant contacts. The degree of node u_1 is equal to the effective size of u_1's local neighbourhood (i.e., three), and the efficiency is therefore one. This places node u_1 in an advantageous brokerage position in which it can bridge three structural holes and intermediate between otherwise disconnected contacts. There is no feasible channel available to the other three nodes for communicating directly and exchanging information with one another. In this sense, node u_1 is needed to secure communication among nodes u_2, u_3, and u_4. However, when the geographic layer is taken into account in addition to the social layer, new links between



(a) Overestimation of brokerage (b) Underestimation of brokerage

Figure 5.1: Brokerage in geo-social networks

node u_1's contacts become apparent: a link between node u_2 and node u_3, and a link between node u_3 and node u_4. This increases the redundancy of node u_1's contacts, thus reducing the effective size and efficiency of u_1's overall (i.e., geo-social) local neighbourhood to 5/3 and 5/9, respectively. An opposite problem of underestimation is illustrated in Fig. 5.1b, where it appears that node u_1 has a fully redundant local neighbourhood in the online social layer, where u_1's contacts (i.e., u_2 and u_3) are connected with each other. In this case, the effective size of u_1's local neighbourhood is one, while the efficiency is 1/2. However, when the geographic layer is taken into account, we find that node u_1 has an additional non-redundant contact, u_4, that has no connection with node u_2 and node u_3. This therefore increases the effective size of node u_1's overall local neighbourhood from one (in the social layer only) to 7/3. Correspondingly, efficiency increases from 1/2 to 7/9. The two examples in Fig. 5.1a and Fig. 5.1b thus clearly suggest that *the analysis of brokerage opportunities can be biased by problems of over- and underestimation when only a single network layer is taken into consideration.*

To overcome this, the notion of brokerage is extended using a multilayer graph, where geographic and social links are both regarded as feasible communication channels that can jointly provide nodes with opportunities for brokering between others. This means that cross-layer triads and triangles must be allowed for. The most straightforward way to achieve this is to calculate the union of the two layers, such that $G^{\alpha \cup \beta} = (V^{\alpha \cup \beta}, E^{\alpha \cup \beta})$, as we have done in Chapter 4. This reduces the problem to a single graph to which the existing measures of brokerage can be applied. Since we are interested in the structure of the resulting combined network, we do not assign different weights to each layer, even though in other problems this may be methodologically appropriate. We will next describe the geographical and social data of which our two layers are composed.

5.1.1 Dataset

The dataset consists of the check-ins and links connecting 37,722 active users of the location-based social network Foursquare in London, UK. This includes 549,797 check-ins, each representing a visit made by a user to a certain venue at a certain time and date. These check-ins have been made to 43,584 venues, and have been posted to Twitter by the users in the period between December 2010 and September 2011, with their respective social networks downloaded at the end of that period. First, a *social network* can be built from the reciprocal Twitter following between all Foursquare users who have shared their check-ins on Twitter. Because the goal is to detect structural holes spanned by users both in the online and in the co-location networks, the focus is only on the Foursquare users of Twitter. The presence of spammers and bots in this dataset is unlikely due to the nature of Foursquare check-ins which require physical presence at a location and are restricted in time, practically infeasible for spammers. Furthermore, we only consider reciprocal social links and spammers are shown to have a very low number of bidirectional links on Twitter [200].

The *co-location network* is built on top of the social network, with the same nodes and using the check-ins posted to Twitter by the Foursquare users. Two nodes in this network are connected if they were co-located, i.e., they happened to be at the same place and at the same time, which reflects the potential for exchanging information offline. The new network is constructed by using the timestamp of users' check-ins to venues, where if two users have checked-in to the same venue within a 1-hour window, a link is placed between them in the co-location network. A link is placed between two users only when they were co-located more than once in order to minimise false positives. In this way, a proxy for offline interaction between Foursquare users is used, in agreement with previous studies of co-location from location-based services [50].

network	$\|V\|$	$\|L\|$	$< k >$	$< C >$
social	36,926	176,164	9	0.15
co-location	8,059	112,367	27	0.51
geo-social	37,722	287,661	15	0.2

Table 5.1: Network properties

Finally, the *geo-social multilayer network* is built by taking the set of links that are produced by the union of the social and co-location networks. The multilayer network is undirected, and links in this network are unweighted. Table 5.1 outlines the following network properties of the two single-layer networks and the combined multilayer network: the number of nodes, V; the number of links, L; the average degree, $< k >$; and the average (local) clustering coefficient [196], $< C >$. The cumulative degree distributions of all three networks are shown in Fig. 5.7.

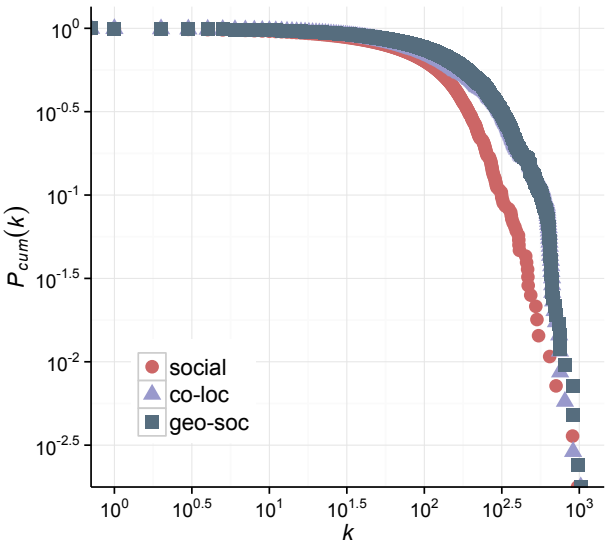


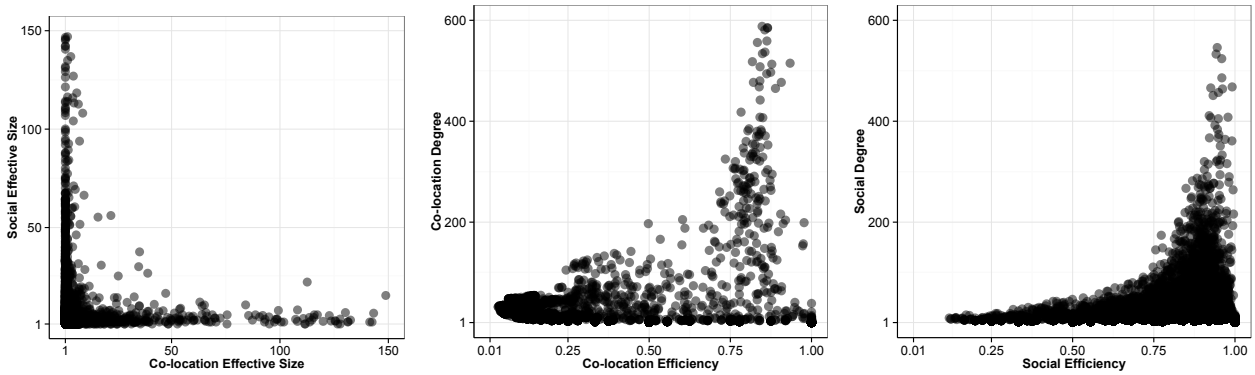
Figure 5.2: Cumulative degree distributions $P_{cum}(k)$ of the single-layer networks and the geo-social network

5.1.2 Evaluation

In this section, the interplay between online social network brokerage and co-location network brokerage in the Foursquare network of users in London is evaluated. To this end, the changes in the effective size and efficiency of users' local neighbourhoods is investigated from one layer to the other, and then the users' brokerage opportunities in the combined multilayer geo-social network is measured.

Social vs geographic brokerage Equation 5.2 is used to capture the brokerage positions of all nodes that belong to both the co-location (geographic) and social network layers. Nodes with degree equal to zero in either layer are therefore removed from the analysis. In Fig. 5.8a, we plot the effective size of nodes in the geographic co-location layer against the nodes' effective size in the social network layer (up to a size of 150 for the sake of visibility). The majority of nodes are associated with a high brokerage potential only in one of the two layers, but not across layers. This suggests that users may seem to intermediate between others when evaluated within a single layer, when in fact their opportunities for brokerage are much fewer when the two layers are combined.

The social network degree and the co-location network degree of nodes bear no correlation. Since a node's opportunity for brokerage greatly depends on the node's degree, it would be unreasonable to expect a correlation between the effective sizes of the node's local neighbourhoods in both layers. This suggests that there is a trade-off between being physically co-located with many others in the co-location network and having many friends online. Correspondingly, there is also a trade-off between brokerage positions online and offline, as indicated by our findings. *Individuals that hold prominent brokerage positions*



(a) Social vs geo effective size (b) Geo degree vs efficiency (c) Social degree vs efficiency

Figure 5.3: Effective size and efficiency of nodes' local neighbourhoods

either in the social or in the geographic layer of the network are not likely to hold an equally prominent position in the combined two-layer network. This also implies that *the assessment of brokerage confined within a single layer in isolation may be positively or negatively biased.*

Figs. 5.3b and 5.11a show the efficiency of each node's neighbourhood as compared to the node's degree. Although the vast majority of users have a low degree and low efficiency, it is clear from the figures that nodes with a degree higher than 200 tend to have an efficiency between 0.75 and 1.00. This suggests that as nodes increase the number of their contacts, they also optimise the efficiency of their local neighbourhoods by including non-redundant contacts, thus expanding the opportunities for brokerage [36]. As shown in Fig. 5.8a, *high-degree nodes appear to be able to secure a higher efficiency of their networks than low-degree nodes.* This finding is consistent with other related studies that have documented an inverse relationship between the degree and the local clustering of nodes [186, 154]. Because nodes with high degrees tend to have a lower clustering coefficient than nodes with low degrees, the former nodes are also expected to leverage on greater brokerage opportunities than the latter ones [102].

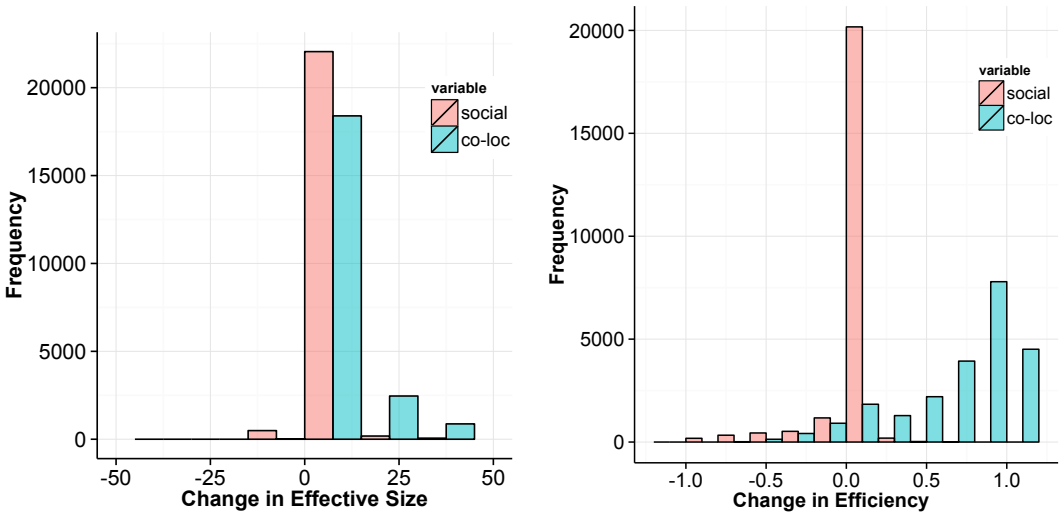
When the effective size of nodes' neighbourhoods is evaluated only in the co-location network, the nodes with high brokerage potential do not correspond to the nodes with high potential in the social network layer. Therefore, a combined geo-social approach to brokerage, which unifies both interaction layers into an integrated source of brokerage opportunities is more suitable.

Geo-social brokerage Geo-social brokerage is here regarded as arising from the union of the co-location network and the online social network. Information can indeed be transmitted both online and offline, and whether individuals can benefit from having access to novel and non-overlapping sources of information can only be investigated by analysing the structural positions individuals hold in the combined online and offline

network. The union of the two layers has an effect on the degree of the individual user. In particular, for each node the degree in the multilayer network cannot be smaller than the lower degree the node has in either single-layer network, and cannot be higher than the sum of the two degrees the node has in the two single-layer networks. On the one hand, the degree in the multilayer network takes on the minimum value when all the contacts a node has in one layer are the same as the contacts in the other layer. On the other hand, it takes on its maximum value when all contacts in both layers are unique. The implications in terms of opportunities for brokerage are straightforward: in the former case, the node has fewer opportunities than in the latter as there are fewer contacts among whom the node can intermediate. However, not all opportunities for brokerage will translate into actual structural holes. This will depend on the variation in links among contacts resulting from the inclusion of an additional layer. When one network layer is combined with another, some of a node's contacts that were unconnected in the former layer may be connected in the latter, thus mitigating the potential of the node to intermediate between contacts. The effective size of the node's neighbourhood in the combined network will ultimately depend on the interplay between variation in number of contacts and variation in number of links between contacts [102].

Fig. 5.4 shows the distribution of change in brokerage induced by the geo-social network with respect to each of the single-layer networks. For each node, this change is here measured as the difference between the effective size and efficiency of the node's neighbourhood in the composite geo-social network and the node's effective size and efficiency in the single-layer network. When a node has a degree (and effective size) equal to zero in either layer, the node's efficiency is set equal to zero in that layer. Fig. 5.4a shows changes in effective size only within the range $(-50, 50)$. As suggested by the figure, there is an improvement of brokerage potential in the geo-social network over brokerage in the co-location network. When the social layer is also accounted for in the analysis of a node's brokerage position, additional structural holes emerge in the node's neighbourhood, thus amplifying the node's opportunities to intermediate among disconnected others. However, while the majority of nodes can also improve their brokerage positions when the co-location layer is added to the social layer, nonetheless there are some who suffer from a decrease in structural holes. Thus, the co-location network may contribute toward increasing the number of a node's unique contacts, but at the same time may also add new links among some of the node's contacts that would appear as unconnected in the social layer. These mixed effects of the geo-social network on brokerage are even more pronounced when assessed in terms of variation in efficiency. As indicated by Fig. 5.4b, while most nodes seem to secure a more efficient neighbourhood when one layer is combined with the other, there are some who suffer from a loss of efficiency, especially when the co-location layer is added to the social one.

Overall, these results suggest that brokerage may be over- or underestimated when assessed in one single layer. *On the one hand, in the online social network layer many users*



(a) Variation in effective size

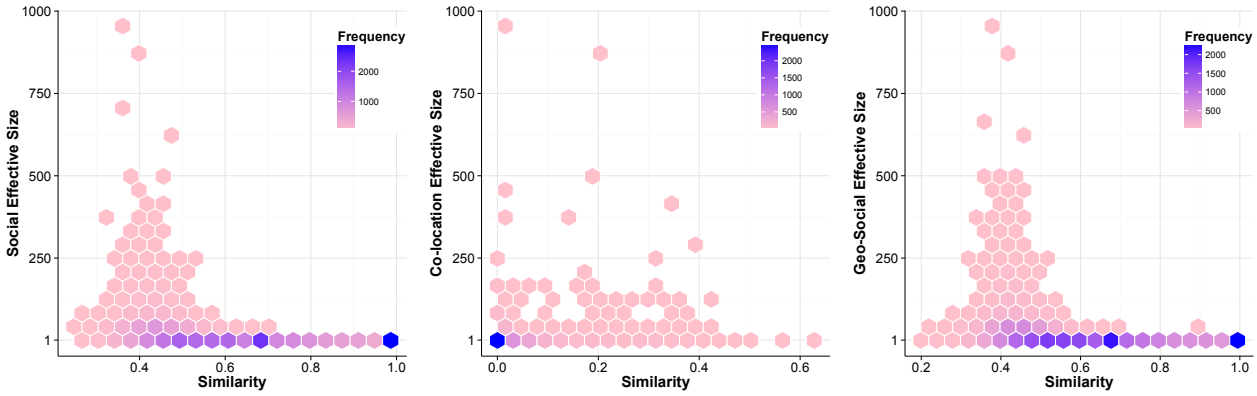
(b) Variation in efficiency

Figure 5.4: Variation in effective size and efficiency between the multilayer geo-social network and single-layer network

may appear to be brokers but their influence and intermediation power may be overrated. On the other, many users who appear to have little brokerage power offline may be active brokers when the offline connections are combined with the online ones.

Neighbourhood heterogeneity One of the main benefits of brokerage lies in the advantage that individuals acquire through novel recombinations of non-redundant information [36, 109, 175]. Brokers are expected to leverage on the diversity of their contacts in order to intensify the competitive advantage they can derive from their structural positions. We test whether in the dataset brokerage positions are associated with heterogeneity of local neighbourhoods. In particular, it is examined whether the association between brokerage and heterogeneity can be detected in each of the single-layer networks, and whether there is a variation in the association when the two layers are combined.

To this end, for each node the average cosine similarity is computed between all unconnected pairs of the node's contacts. In turn, similarity between unconnected contacts is assessed by using the frequency distribution vector of the categories of places that each user has visited. This vector is assumed to be representative of the user's personal preference for categories of places visited. There are nine top-level categories of places in Foursquare: "Professional & Other Places", "Shops & Services", "Travel & Transport", "Food", "Nightlife Spots", "Arts & Entertainment", "Colleges & Universities", "Outdoors & Recreation", and "Residences". A nine-cell numeric vector is built, in which each cell is representative of a category and contains the frequency of visits made to that category by each user within the period covered by the dataset. For each node, and for each pair of users in the node's neighbourhood, the two corresponding vectors are compared, and an average score of similarity associated with the focal node's neighbourhood is obtained.



(a) Social network (b) Co-location network (c) Geo-social network

Figure 5.5: Effective size of nodes in different layers as compared to their average neighbourhood cosine similarity

Fig. 5.5 shows, for each node, the relationship between the effective size of the node's neighbourhood and the node's average neighbourhood similarity. In agreement with the literature [36, 79], Fig. 5.5a suggests that users with many opportunities for brokerage in their social networks tend to belong to a heterogeneous neighbourhood, while the majority of users that are embedded in socially cohesive networks (i.e., with low effective size) tend to be connected with similar others. Findings on the co-location layer are mixed, even though users with the highest brokerage potential are still associated with relatively high diversity in their local neighbourhood (Fig. 5.5b). Finally, when both layers are combined into the geo-social network, neighbourhood heterogeneity is associated with brokerage opportunities (Fig. 5.5c). All users with an effective size of more than 250 have an average neighbourhood similarity of no more than 0.6. Thus, the geo-social network retains and reinforces the pattern observed in the other two layers.

In conclusion, the analysis of multilayer brokerage strengthens the study of social capital by uncovering sources of information benefits that would otherwise remain hidden in a single-layer network. For instance, findings indicate that the potential for novel recombinations of non-overlapping information would remain undetected if only the co-location layer were taken into account. Only when this layer is combined with the online social network layer can the heterogeneity of a node's unconnected contacts be fully captured and the value the node can extract from structural holes be properly assessed.

Having examined the brokerage power of individuals from a multilayer perspective, a follow-up question on the brokerage power of places emerges. Urban places are designed to bring together people and foster social interactions, so can places be brokers too? In the remainder of this chapter we will attempt to answer this question through a broader analysis of urban social diversity and what it means for urban development and location-based applications.

5.2 The Social Diversity of Places

People and places are interconnected in an organic way [17], and more intensely so in the urban context. With more than half of the world's population living in urban areas,[1] understanding how *human mobility enhances the social diversity of places* is important for urban planners and system designers alike. While the fundamental role of urban geography in human interactions, relationships and social capital is relatively well understood [44, 63, 140], the role of people in the success of places has been empirically understudied. Nevertheless, much of the success of cities can be attributed to their multiculturalism and the synergy of diverse attitudes within a relatively small geography that is brought on by its inhabitants and measurable through their human mobility network and social network properties [71].

Urban activist Jane Jacobs wrote "[Cities] differ from towns and suburbs in basic ways, and one of these is that cities are, by definition, full of strangers" [92]. In this section the places that bring together strangers among other types of urban social diversity is explored and the relationship with the prosperity of an area in an interconnected model of people and places is measured. In social networks, the diversity of one's social ties is associated with the amount of social capital they have at their disposal as demonstrated in the previous section of this chapter. Social ties can be classified as bridging or bonding where bonding ties are those within homogeneous groups while bridging ties are those which transcend groups and are associated with diversity [150]. *However, it is not yet understood how the diversity of the social network of visitors defines the social role and affluence of a place.*

5.2.1 The Interconnected Geo-Social Network

Many real-world systems can be represented through a number of unique yet interconnected networks. One such system results from the interaction between geography and people. Although the properties of social and geographic networks have long been studied independently, such view does not consider the dynamics between the two. Here a model of interconnected geo-social networks is presented, where projections of one carry rich information about the other.

The spatial graph $G_L = (V_L, E_L)$ has a set of nodes $l \in V_L$ which are geographical locations and can be described by a set of coordinates, and a set of edges E_L that can be described in terms of the user transitions between them with a weight equal to their number. The neighbourhood of a location l can be denoted as $N_L^h(l)$ and includes all its adjacent associated locations in terms of transitions up to h hops.

[1] United Nations, 2014 Revision of World Urbanisation Prospects. `http://esa.un.org`

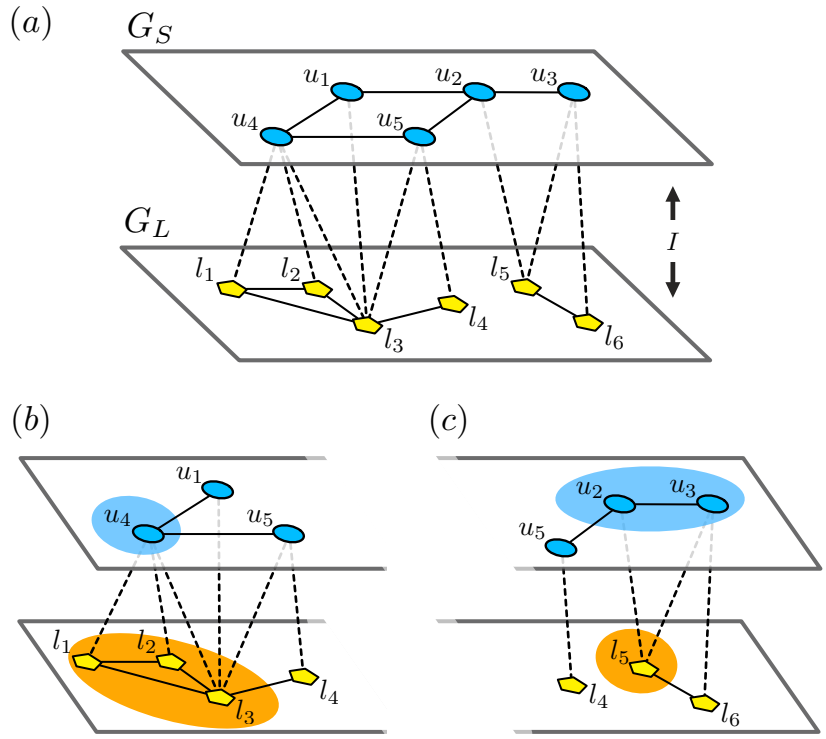


Figure 5.6: Panel (a): Interconnected network model, where G^L and G^S are composed by different entities coupled through interlayer edges I. Panel (b) shows the place neighbourhood of user u_4, as indicated by the shaded place nodes in the lower layer. Panel (c) illustrates the social neighbourhood of visitors to place l_5, indicated by the shaded user nodes in the upper social layer.

The social graph $G_S = (V_S, E_S)$ includes nodes which represent users (denoted as V_S) and undirected edges which are the friendship relationships between them (denoted as E_S). The neighbourhood of a user u can be denoted as $N_S^h(u)$ and includes its contacts up to h hops in the social network. The interconnected network $G_M = (G_S, G_L, I)$ contains the geographic and social graph layers along with the interlayer edge set I. We associate a weight $w_{u,l}$ to each edge which equals the number of visits (check-ins) that a user u has made to a location l.

The social network of the individuals linked to a location l at distance h can be described as its social neighbourhood, denoted by $N_S^h(l)$. For example, the 1-hop social neighborhood of location l would be composed of the individuals that visit location l, the 2-hop social neighborhood would be composed of the individuals that visit location l and their friends and so on. On the other hand, the place network of an individual u at hop distance h can be denoted by $N_L^h(u)$. It is a subgraph of the spatial network layer, where each place $l \in N_L^h(u)$ is at distance less or equal than h from user u. In the case of $h = 1$, it is the set of places user u has visited. For $h = 2$ it contains the places visited by u and the places connected to those visited by the user in the place network and so on.

Fig. 5.6 illustrates the interconnected geo-social network model in Panel (a). Locations are connected based on their common visitors, and interlayer edges I represent visits made

by users to locations. Panel (b) illustrates the place neighbourhood of a user node u_4 as a projection on to layer G_L, while Panel (c) shows the reverse projection of a location l_5 on to the social network G_S. Both of these projections are used to construct measures of urban social diversity in the following section.

5.2.2 Urban Social Diversity Measures

In this section, four measures of the social diversity associated with a place are defined through its social network of visitors. *Brokerage* relates to the potential of a place to bring together strangers as opposed to friends, *serendipity* measures the probability that the set of visitors happened to visit that place, while *entropy* measures the diversity of visits with respect to visitors. The *diversity* of visitors themselves is also measured by comparing their characteristics in terms of venues.

Brokerage. As explored in the previous section, the brokerage potential of a person expresses his or her ability to connect otherwise disconnected others. Extensively described by Burt, it measures the extent to which an individual's ego network is non-redudntant, which in turn reflects the individual's potential of brokering between otherwise disconnected contacts [37]. Within the context of geography, a place can possess brokerage potential with respect to the social network of its visitors, if it can bring together otherwise disconnected individuals in physical space. Based on the interpretation by Borgatti [23], the brokerage potential B of node l at distance h can be expressed as:

$$B(l) = |N_S^h(l)| - \frac{\sum_{u,v \in N_S^h(l)} e_{u,v}}{|N_S^h(l)|} \tag{5.2}$$

where we subtract the redundant portion of a network, equivalent to the connectedness (average number of edges $e_{u,v}$) in the social neighbourhood of l, from its size $|N_S^h(l)|$. This can then be normalised by $|N_S^h(l)|$, resulting in the fraction of non-redundant contacts of l's social neighbourhood: $B(l)/|N_S^h(l)|$. In this work we use a hop parameter $h = 2$, which enables us to capture second-hand redundancy resulting from connections among friends of visitors. If all visitors to a place are connected, the place has no brokerage power ($B(l) = 0$), whereas if none of the visitors are connected, the place has high brokerage power which results in a brokerage value of 1.

Serendipity. The serendipity of a place is the extent to which it can induce chance encounters between its visitors. This can be measured as the average probability of an edge $w_{u,l}$, given the network of places u has visited prior to venue l. This expresses the idea that all visitors to l have arrived there with a certain probability based on the network

of places they have visited in the past. The lower the probability, the higher serendipity a place can provide. More formally, we can define the serendipity D of a place l as:

$$D(l) = 1 - \frac{\sum_{u \in N_S^h(l)} p_l^t(u)}{|N_S^h(l)|} \tag{5.3}$$

where:

$$p_l^t(u) = \frac{\sum_{v \in N_L^h(u)^{<t}} w_{v,l}}{\sum_{v \in N_L^h(l)^{<t}} w_{v,l}} \tag{5.4}$$

is the probability of user u checking into place l based on their place neighbourhood $N_L^h(u)^{<t}$, and t represents the first check-in to venue l made by user u. The probability is measured as the sum of weights of the number of venues v with edges to l visited by u prior to time t, over the weighted degree of l in the spatial network. The average probability of a user u visiting location l provides a measure of what role chance plays in the composition of the social neighbourhood of l. Places with a higher serendipity value are more likely to induce chance encounters since the composition of their visitors is more unexpected.

Entropy. The entropy of a place describes the extent to which it is diverse with respect to visits. Its value can be measured as the Shannon entropy of a location:

$$H(l) = -\sum_{u \in N_S^h(l)} p_l(u)\ log p_l(u) \tag{5.5}$$

where $p_l(u)$ is the probability that a given check-in in place l is made by user u. This measure is applied in a similar way to the authors in [50], where it is used to quantify the diversity of visitors to a location. This is adapted to the definition of the the social neighbourhood of a place, where places with highly entropic neighbourhoods are frequented by many diverse visitors and vice versa.

Homogeneity Another important measure of the social diversity of a place is the extent to which its visitors are homogeneous in their characteristics. The mean cosine similarity between the place preferences of all pairs of visitors to a particular location can be used to measure its overall social homogeneity as:

$$S(l) = \frac{\sum_{u,v \in N_S^h(l)} sim(u,v)}{|N_S^h(l)|(|N_S^h(l)| - 1)} \tag{5.6}$$

where $sim(u,v)$ is the cosine similarity of the frequency vectors of the visits to locations of a given category of user u and user v respectively. There are nine top categories in

Foursquare for which we build a frequency vector for each user and then compare in a pairwise manner for all visitors of a venue. These categories are further described in the Dataset Section 5.2.3. We derive homogeneity between users in a similar way to the authors in [48] in that we consider the cosine similarity of user activity.

This value is between 0 and 1 and indicates the extent to which the mobility patterns of a pair of users in terms of categorical venue visits are the same (1) or completely different (0). By averaging these values across the social neighbourhood of a place, we can derive an estimate of the homogeneity of its visitors in terms of venue preferences. We will describe the data which forms our interconnected network model next.

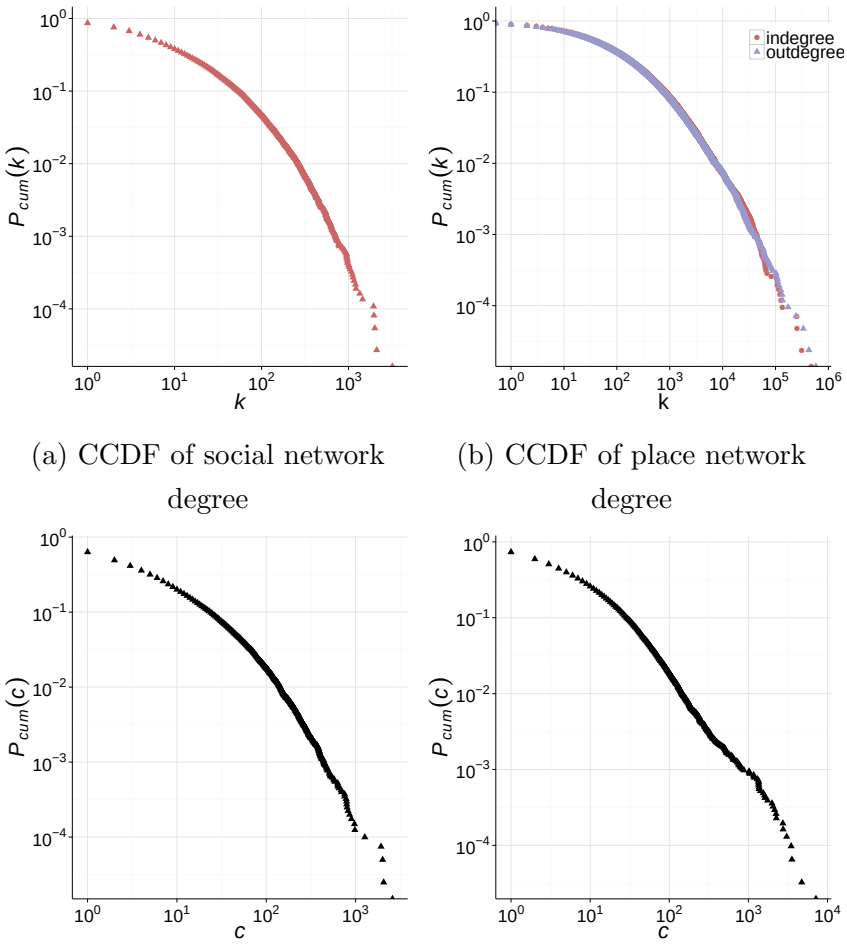
5.2.3 Dataset

The dataset on which evaluation was performed is the same as in the previous section but applied in a novel way to reflect the interconnected nature of people and places. This section describes how this data fits the interconnected geo-social network model.

Online Social Network. Similarly to the previous section, an undirected *social network* is built from the directed Twitter network where user u follows v and v follows u back for all Foursquare users who have shared their check-ins on Twitter and we therefore refer to two users with a reciprocal edge as *friends* in this analysis. The undirected social network of London users consists of $432,929$ unweighted reciprocal links between $36,926$ users. In Fig. 5.7a we can observe the cumulative degree distribution of the social network, having a long-tail with a minority of very well connected users and the majority of users with less than 100 friends.

Place Network. The spatial network of places is constructed by the transition flows of users going between locations. These spatial locations are referred to as *places* and their network as a *place network*. If a user has transitioned between two places in their history of check-ins, we draw an edge between them. The weight of the edge is proportional to the number of transitions made by all users between two places and edges are directed. In total, there are $3,151,741$ directed edges between the $42,080$ venues. The degree distribution is shown in Fig. 5.7b.

Each venue in Foursquare is also characterised by a lower level category such as *coffee shop* and a higher level category such as *Food*. There are nine top-level categories: *Arts & Entertainment, Colleges & Universities, Food, Nightlife Spots, Outdoors & Recreation, Professional & Other Places, Residences, Shops & Services, Travel & Transport*, which we refer to for short in this work as *Arts, Study, Food, Nightlife, Outdoors, Professional, Residences, Shops* and *Travel*. In addition, each venue falls within a geographic administrative boundary called a borough. Each borough consists of wards, which are sectioned



(a) CCDF of social network degree

(b) CCDF of place network degree

(c) CCDF of check-ins per user

(d) CCDF of check-ins per venue

Figure 5.7: Cumulative degree distributions of the social and place graphs as well as of interlayer edges (check-ins).

by population density and the natural landscape of the city. We use categories and geographical boundaries to distinguish between measurement effects in our results and government statistics of deprivation.

Geo-social Interaction. The interaction between people and places in the dataset is represented by Foursquare check-ins. There are more than half a million check-ins that we have recorded in the area of London over a little less than a year. Figs. 5.7c and 5.7d plot the distribution of check-ins per user and check-ins per venue respectively. A very small fraction of users have made an exceptional number of check-ins over the time period and similarly most venues have a low number of check-ins with the exception of some highly popular venues. Heathrow Airport is the most popular venue in London with over 10,000 check-ins in our dataset.

Index of Multiple Deprivation. To quantify socioeconomic conditions within regions of London the Index of Multiple Deprivation (IMD) is used, an official statistical exercise

conducted by the UK Department of Communities and Local Government to assess the relative prosperity of neighbourhoods across England. The overall IMD for an area is a composite of seven deprivation indices. In particular, a neighbourhood is assessed according to the following domains: deprivation relating to low income (*Income*); deprivation due to lack of employment among working-age inhabitants (*Employment*); lack of education and skills among young persons and adults (*Education*); impaired quality of life due to ill health and disability (*Health*); risks of crime at a local level (*Crime*); limited provision of local services and lack of access to affordable housing (*Housing*); deprivation relating to the local environment, including quality of housing and air quality (*Living Environment*). The composite IMD and seven domain indices for each neighbourhood are publicly available[2], and provide a rich source of curated socioeconomic indicators across London. The higher the score of an index, the more deprived the neighbourhood. In this evaluation, the indices released with the two most recent reports are considered (2010 and 2015).

5.2.4 The Social Role of Places

The results of the evaluation of urban social diversity measurements are presented in this section. First, a distinction between the *bridging* (bringing together strangers) and *bonding* (bringing together friends) qualities on a per venue basis and between categories is made. The *diversity of visitors* to those venues in terms of their characteristics is then explored. Ultimately, these observations are related to *neighbourhood deprivation* and the differences between central and peripheral boroughs of London with regards to social diversity are reported.

The Brokerage Role of Places. One of the fundamental social roles of places is to bring people together. Just like people in social networks, some places can act as *bonding hubs*, bringing together friends to socialise and interact with each other, while others are more likely to gather strangers and therefore act as *bridging hubs*, bringing together otherwise disconnected individuals. The bridging or bonding role of a Foursquare venue is here measured as its brokerage $B(l)$ using Equation 5.2.

The role of a place to either bring together friends or strangers can be dependent on its type. Fig. 5.8a shows the distribution of brokerage values across categories. In the box-and-whisker plot, the distribution of brokerage is split into quartiles. Each box represents the mid-quartile range with the black line in the middle being the median of the distribution. In a megacity such as London, it is unsurprising that most locations are frequented by many diverse individuals who do not know each other. There are, however, notable

[2]English Indices of Deprivation, 2015.
`https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015`

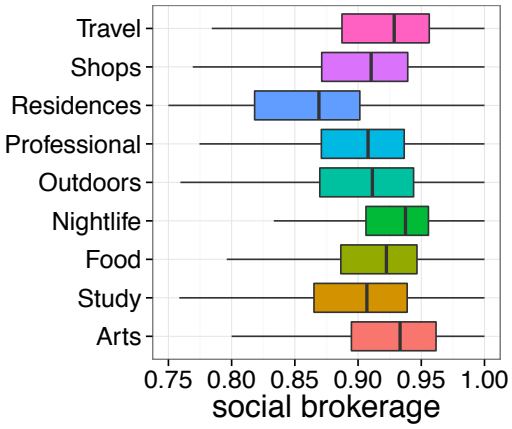
variations between some of the categories. *Residences* tend to be bonding hubs with 50% of values below the distribution median value of 0.87, followed by *Study, Professional, Shops* and *Outdoors* categories where people are more likely to be with friends. Places with relatively high brokerage are in the *Arts, Nightlife* and *Travel* categories where most places in these categories have a bridging role in bringing together strangers.

While the structure of the social network of visitors to a place can determine its brokerage role, serendipity further explains how probable its composition of strangers or friends is and to what extent it can foster encounters, which may lead to new social interactions rather than pre-determined ones. It measures the average probability that a person visited the location given their prior history of locations (Equation 2). Fig. 5.8b plots brokerage against mean serendipity. While serendipity varies for low values of brokerage, the relationship between the two is positively strong for higher values of brokerage. This suggests that bonding hubs which are more socially cohesive may have a lower ability to induce chance encounters, while high bridging places will have high serendipity.

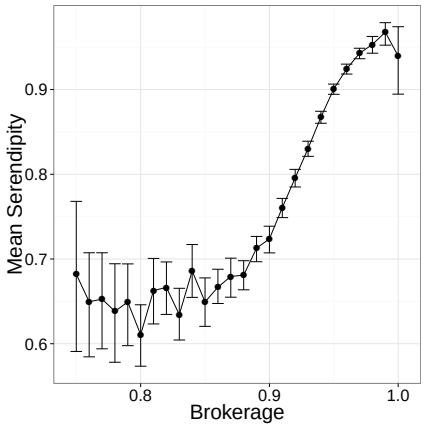
We take a closer look at the sub-categories of places and their brokerage and serendipity roles in Table 5.2 where the top bridging (highest value) and bonding (lowest value) types of places and their serendipity values are listed. Firstly, within the *Arts* category there is a clear distinction between the types of places that bridge which seem to be associated with public spaces, while bonding places tend to be predominantly sports/team oriented. Similarly in the *Study* category it is interesting to observe that academic buildings tend to be bridging, while specific departments and classrooms are places that friends tend to have visited together. Within the *Food* category, interestingly fast foods appear as having a greater bonding role than international cuisines such as Australian and German. In the case of *Nightlife*, however, more generic sub-categories such as *Bar* or *Pub* have greater bridging roles than more specific nightlife venues such as *Hookah Bar* or *Strip Club*, which tend to bring together friends. This table is provided only for intra-category comparison as similarly to the distributions shown in Fig. 5.8a, the brokerage values are not entirely distinct across categories and we use the overall brokerage in all subsequent analysis.

Similar to the *Arts* category, the *Outdoors* category seems to be split between sports-related activities and public spaces. In contrast to *Arts*, however, here sports are being played rather than watched and brokerage values are generally lower possibly because the team activity is more bonding than the viewing activity. In the *Professional* category we observe some public service places such as *Courthouse* and *Hospital* to be bridging hubs, where mostly strangers visit with relatively high serendipity. On the other hand, *Doctor's Office* and *Elementary Schools* are venues where more friends check-in, similar to religious venues such as *Mosques* and *Synagogues*. Because visitors to such venues are likely to be locals, it is understandable that they may know each other and be regulars at these venues.

The *Residences* category has only four sub-categories, which all have relatively low bro-



(a) Brokerage of urban places by category



(b) Brokerage vs mean serendipity

Figure 5.8: Brokerage per category and brokerage vs. serendipity. Outliers below brokerage values of 0.75 (3% of values) were excluded for readability.

Category	*Top Bridging*	$< B >$	$< D >$	*Top Bonding*	$< B >$	$< D >$
	Aquarium	0.98	0.98	Basketball	0.85	0.72
	Art Museum	0.95	0.84	Billiards	0.88	0.85
Arts	Opera House	0.96	0.97	Football	0.87	0.52
	Cricket	0.94	0.75	Track	0.87	0.74
	Theatre	0.94	0.87	Water Park	0.9	0.78
	Auditorium	0.92	0.9	Classroom	0.86	0.83
	University	0.91	0.82	Communications	0.89	0.93
Study	Lab	0.91	0.88	Engineering	0.85	0.56
	Rec Center	0.88	0.84	Math	0.69	0.45
	Bookstore	0.9	0.79	Medical School	0.84	0.7
	South American	0.92	0.78	Eastern European	0.88	0.75
	Scandinavian	0.94	0.83	Wings	0.8	0.59
Food	German	0.95	0.91	Indian	0.88	0.71
	Dumplings	0.93	0.88	Friend Chicken	0.87	0.62
	Australian	0.95	0.72	Felafel	0.89	0.76
	Lounge	0.93	0.81	Hookah Bar	0.88	0.83
	Gay Bar	0.92	0.86	Strip Club	0.89	0.77
Nightlife	Pub	0.93	0.85	Hotel Bar	0.89	0.7
	Cocktail	0.92	0.84	Dive Bar	0.87	0.75
	Bar	0.92	0.83	Whiskey Bar	0.88	0.83
	Bridge	0.89	0.87	Athletics & Sports	0.85	0.6
	Neighbourhood	0.9	0.83	Baseball Field	0.75	0.72
Outdoors	River	0.9	0.76	Campground	0.85	0.85
	Park	0.9	0.79	Vineyard	0.69	0.63
	Cemetery	0.9	0.8	Soccer Field	0.82	0.77
	Hospital	0.91	0.87	Emergency Room	0.81	0.82
	Landmark	0.91	0.81	Synagogue	0.83	0.79
Professional	Courthouse	0.9	0.8	Mosque	0.87	0.63
	Convention Centre	0.91	0.77	Elementary School	0.68	0.88
	Animal Shelter	0.93	0.93	Doctor's Office	0.84	0.65
	Residence	0.84	0.46	Housing Development	0.83	0.52
Residences	Apartment Building	0.86	0.75	Home	0.82	0.69
	Photography Lab	0.96	0.9	Yoga Studio	0.88	0.79
	Antiques	0.92	0.82	Laundry	0.72	0.83
Shops	Mall	0.93	0.9	Video Store	0.72	0.71
	Gift Shop	0.93	0.74	Gaming Cafe	0.86	0.51
	Travel Agency	0.95	0.3	Tanning Salon	0.84	0.6
	Motel	0.91	0.81	Resort	0.88	0.69
	Pier	0.94	0.84	B&B	0.87	0.61
Travel	Subway	0.95	0.93	Taxi	0.82	0.41
	Light Rail	0.93	0.88	Plane	0.86	0.63
	Platform	0.94	0.91	Bus	0.86	0.72

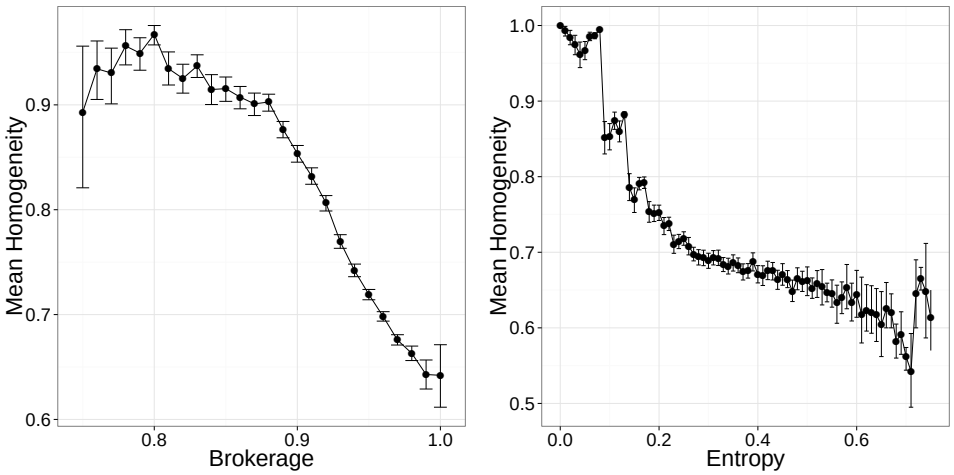
Table 5.2: Top bridging and bonding subcategories by category where $< B >$ is the average brokerage value for the subcategory, while $< D >$ is its serendipity value.

kerage values, but those which play more of a bonding role are *Home* and *Housing Development*, while *Residence* and *Housing Development* seem to have higher bridging roles. It is interesting to also consider the serendipity values for these sub-categories where *Residence* and *Housing Development* are more predictable because visitors have arrived there with higher probability than in the other two sub-categories. *Shops* have a more unexpected mixture of top bridging places except for the *Mall* sub-category where it is likely that many strangers cross paths. Bonding shops are mainly hobby-related and further analysis of this category can reveal potentially intriguing business insight. Finally, the *Travel* category has a definitive split between transportation (bridging), where people tend to commute and travel, a bonding role where people tend to journey with friends. There is, however, the notable exception of *Motels*, which have a bridging role and *Buses*, which play a more bonding role. The brokerage role of places and their serendipity are related but differ across categories. In this section, we have taken a close look at these differences and will compare them to the diversity of visitors and their characteristics in the following section.

Visitor Diversity. While bringing together strangers can ultimately lead to new social exchanges, these might not be truly diverse if the population of visitors is homogeneous. We measure the social composition of a place's visitors with respect to their venue preferences derived from mobility patterns through check-ins. For every user who has checked into a venue, we construct a vector of the frequency of their visits to the nine top-level venue categories. By averaging the cosine similarity of vectors between all pairs of visitors to a place, we obtain the *homogeneity* of a place as per Equation 5. In mobility studies entropy has been used to observe the geographical diversity of contacts that people have and the probability of co-location at diverse venues [63, 164]. However, *a location with high entropy is not necessarily one with visitors with diverse characteristics.*

In Fig. 5.9 the mean homogeneity per venue to entropy and brokerage is compared. In both graphs, we can see that homogeneity decreases as brokerage and entropy increase. This relationship implies that the more diverse the visitors to a place are in terms of their composition and social network connectivity, the more characteristically diverse they are too as measured by homogeneity. Fig. 5.10 further shows the distribution of homogeneity over different values of entropy across categories. The strongest relationships are present in the *Food* and *Nightlife* categories where the highest frequency of values are those with low entropy and high similarity, gradually becoming more spread out as entropy grows. Most other categories exhibit similar patterns, with the notable exception of the *Residences* category for which entropy and brokerage are in general low, yet the trend of decreasing homogeneity with entropy is still present. Overall, *venues, which exhibit high diversity in terms of entropy and brokerage also have a less homogeneous composition of visitors.*

The diversity measures that are introduced here take into account different projections of subgraphs in the interconnected network such as the place network of users in the



(a) Brokerage vs mean homogeneity (b) Entropy vs mean homogeneity

Figure 5.9: Visitor diversity with respect to brokerage and entropy.

serendipity measure and the social network of a venue's visitors in the other measures. Some measures like entropy and serendipity are based on probability while brokerage is purely structural. These differences could be of interest to the designers and developers of mobile systems in improving recommendations for location-based services. In the following section, we focus further on the urban development applications of our research in terms of identifying deprivation and neighbourhoods undergoing gentrification.

5.2.5 Diversity and Urban Deprivation

In social network analysis it has been suggested that individuals who act as brokers often have higher social capital at their disposal [37]. In terms of human geography, it is known that those who communicate with others in geographically diverse regions tend to come from less deprived areas [63]. However, *it is not yet understood whether within the urban context, places which act as bridging hubs are in fact within more well-off areas.* In this section, we observe the geographical distribution of the four diversity measures across the 32 London boroughs. To address the above question, a correlation analysis is performed on the diversity measures on a per-borough basis with the eight indicators of socioeconomic wellbeing included in the IMD.

Fig. 5.11 shows the mean values of diversity measures aggregated per area. What becomes immediately apparent is that there is a clear distinction to be made between inner and outer boroughs in terms of diversity. Figs. 5.11a to 5.11c show notably higher diversity within central boroughs and lower diversity in the periphery. This suggests that *the social diversity of places is highly dependent on geographic factors.* Venues within the central areas of London have higher entropy (Fig. 5.11b) and bring together more strangers (Fig. 5.11a) who are less homogeneous (Fig. 5.11c). Geography, however, is not a big factor for serendipity (Fig. 5.11d) and high serendipity can be high or low in both inner

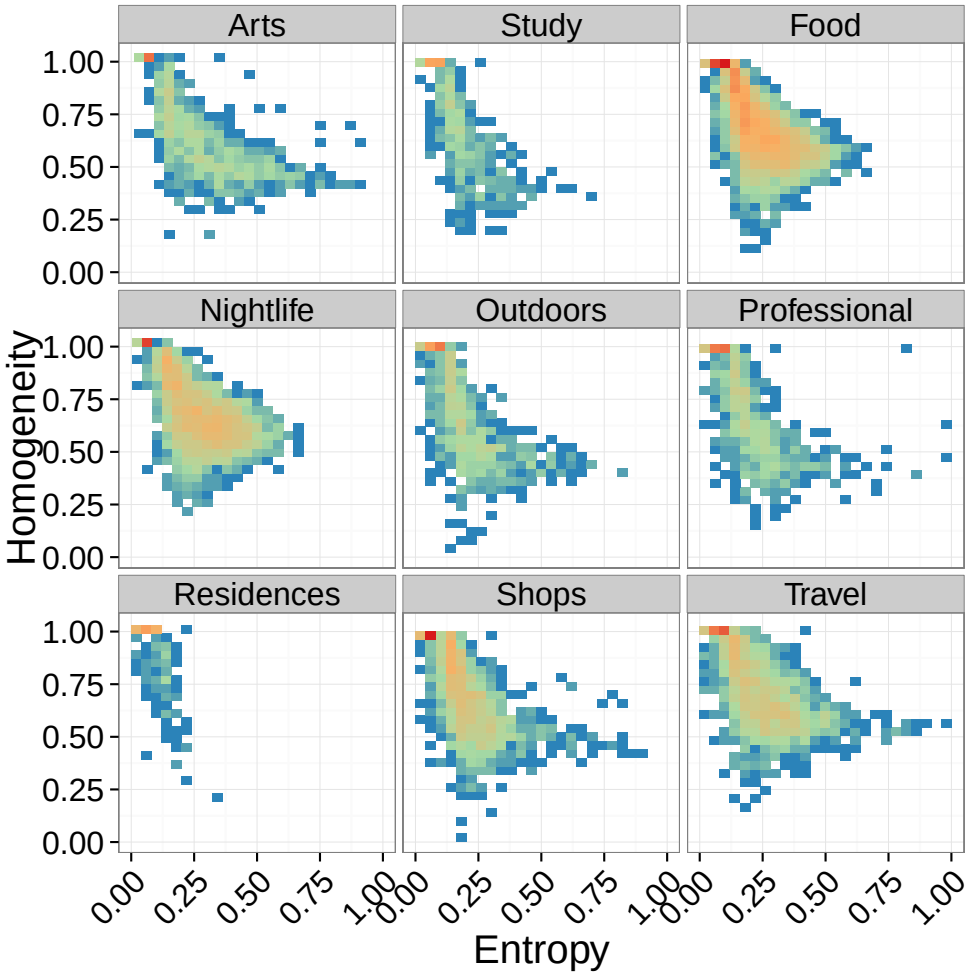
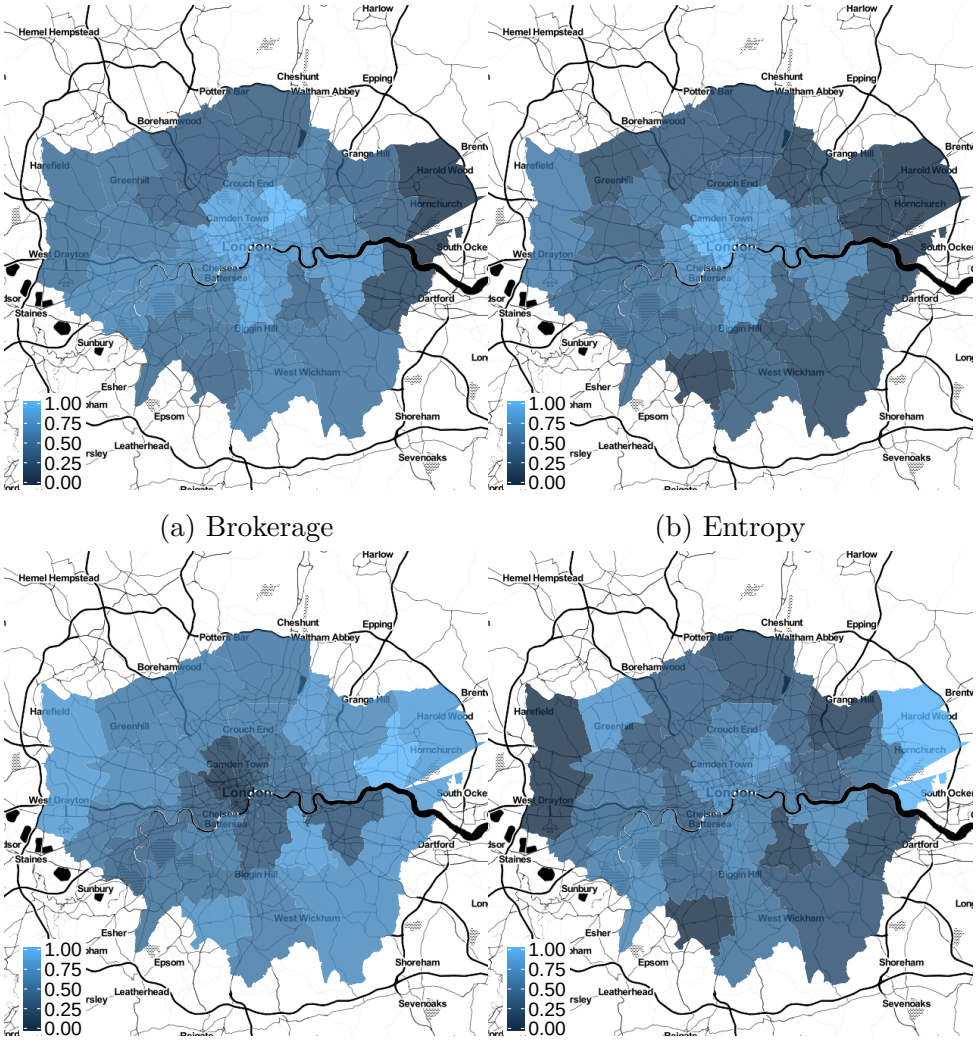


Figure 5.10: Entropy vs homogeneity distribution per category. Colour gradient reflects the frequency of observation with red being high and blue being low.

and outer boroughs. Because the measures are normalised on a per user basis this does not reflect the general popularity of an area on Foursquare.

The ranked correlation between the four diversity measures and deprivation is studied next. While, with regards to social networks, studies have found a positive relationship between diversity and prosperity, in our analysis of the social diversity properties of places, however, *we find that there is a positive relationship between deprivation and diversity in London*. In Fig. 5.12, we note that brokerage has the strongest relationship with deprivation indicators, especially the Living Environment Deprivation score (ρ = 0.71). This sub-domain of the IMD is made up of the following indicators: housing in poor condition, housing without central heating, number of road traffic accidents involving injuring pedestrians or cyclists and low air quality. This sub-domain is especially high for central boroughs where there is more pollution, traffic and strains on housing and health services as indicated also by the relationship with Housing and Health sub-domains ($\rho = 0.44$ and $\rho = 0.4$ respectively). As an example, Westminster, one of the most popular and visited boroughs in central London (home to Buckingham Palace), is also amongst the



(a) Brokerage (b) Entropy

(c) Homogeneity (d) Serendipity

Figure 5.11: Geographic distribution of diversity metrics per London borough.

more deprived ones due to its lack of services, which could be a result of its low number of residents. Correlations with these sub-domains also explain the overall correlation with the IMD itself (ρ = 0.4), however, this is not to suggest that all socially diverse neighbourhoods are deprived in all respects (they may still have high income for example).

Fig. 5.13 plots the relationship between brokerage diversity and IMD for the individual borough. We find that there are in fact quite a few cases where the intuitive relationship between diversity exists. City of London and Kensington and Chelsea, presently two of the wealthiest areas in London, have high diversity in terms of brokerage and low deprivation, while Brent and Lewisham have low diversity and high deprivation. *So why do some places with low diversity have low deprivation, and others with high diversity have high deprivation?* Low brokerage signifies high social cohesion and places which tend to bring together friends have a bonding function. In the context of cities, where millions of people come in contact every day, diversity varies between high and extremely high as we have seen in the previous sections. Our results place emphasis on the value of bonding

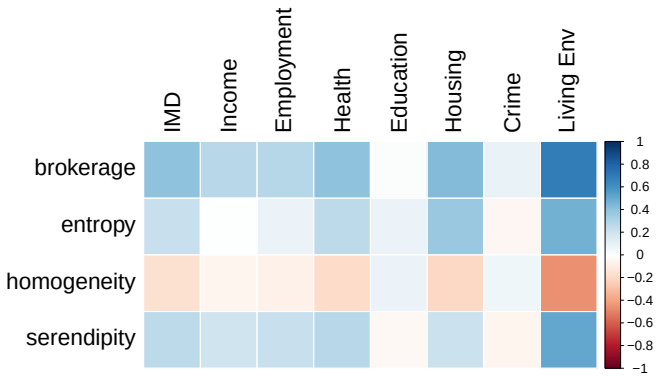


Figure 5.12: Spearman rank correlation matrix of diversity metrics and indicators of deprivation for London boroughs. All correlations are statistically significant with p-values <0.05.

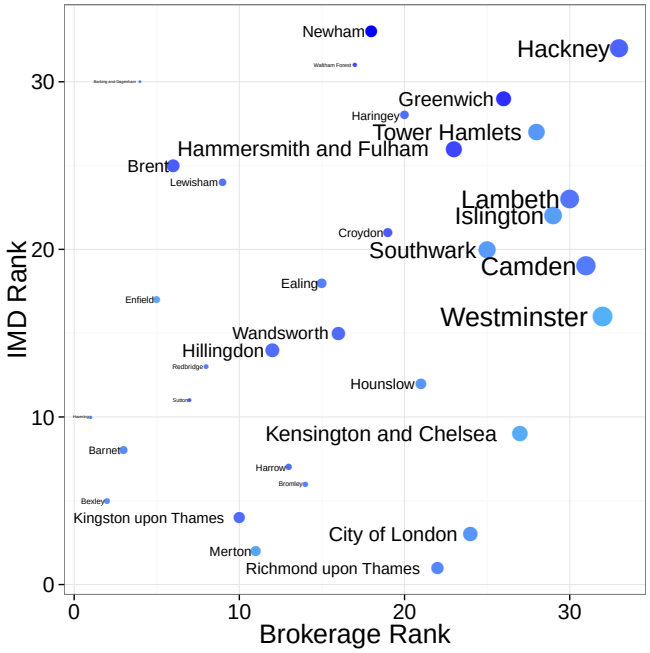


Figure 5.13: Scatterplot of IMD vs Brokerage rank. Size of each node is proportional to its entropy value and colour represents the change in IMD between 2010 and 2015 from low change (light blue) to high change (dark blue).

places within megacities such as London by showing that *more socially cohesive venues are within more well-off areas in terms of overall IMD.*

Now we turn to areas where venues have high diversity and high brokerage, which may seem counter-intuitive at first. In Fig. 5.13 the size of each node is proportional to its entropy and its colour to the difference between the IMD rank for 2010 and the recently released ranks for 2015 (darker blue nodes are associated with larger differences). A further look at the boroughs at the top right corner reveals that they all have high entropy - so highly diverse in terms of visitors (similar patterns were observed for homogeneity and serendipity as previously shown in Fig. 5.11). These areas also experienced some of the most dramatic improvements of IMD between 2010-2015. The London borough of

Hackney, which has the highest brokerage value is the second most deprived but also one of the boroughs with the highest improvement in IMD over the five-year period. It has also been reported that Hackney is currently experiencing the most intense gentrification with fast-rising house prices well above the London average, fast-decreasing crime rate and an exceptionally diverse population [179].

Gentrification is a form of urban migration of affluent citizens to a relatively deprived area, boosting economic development, lowering crime and triggering the renewal of infrastructure and buildings in the area. These benefits come at a cost to the original population of the deprived area, dramatically increasing the cost of living and driving them out of the area, eventually shifting the ethnic profile and characteristics of the neighbourhood which has been found to decrease social capital [104]. In addition to Hackney, Tower Hamlets, Greenwich, Hammersmith and Lambeth among other boroughs in the right-hand corner of the scatterplot are also considered to be undergoing gentrification, suggesting an interesting potential for quantifying this process at the intersection of high diversity and deprivation through social media. Predicting gentrification of neighbourhoods could help local governments and policymakers improve urban development plans and alleviate the negative effects while harvesting economic growth.

5.3 Related Work

Social Capital and Network Diversity. Social capital refers to the value embedded in social networks in the form of connections which can potentially offer support (strong ties) and opportunities (weak ties), and more generally yield expected returns in the marketplace, including the community, the economic, financial, and political markets [109]. Individuals who maintain high social capital in the form of information brokerage are known to have more diverse neighbourhoods both in terms of novelty [36] and geographic dispersion [63], which provides them with advantageous resources as compared to their peers. The competitive advantage of actors in social networks has been formalised as a function of the structural holes that the actors span between otherwise disconnected pairs of others. Spanning holes provides actors with access to negotiation and mediation power [36]. Brokerage and structural holes have been studied in a variety of contexts including online social networks [113]. However, relatively little work has so far examined the interplay between online and offline relationships and their combined contribution to the generation of social capital.

By placing emphasis on the correlation between tie weakness and the flow of information, Granovetter set the stage for a conception of social capital based on discontinuities in social structures and brokerage opportunities. The idea that social capital can originate from brokerage opportunities stemming from structural holes has been further explored by Burt, especially in organisational domains [36]. Burt defines a structural hole as the

"separation between non-redundant contacts", "a buffer" that enables the two contacts to "provide network benefits that are in some degree additive rather than overlapping" [36, p.18]. Burt further identifies two sources of the social capital originating from structural holes: information benefits and control benefits. The less an actor's contacts are already connected with each other, the more likely the actor is to obtain non-redundant information and to reap control advantages by trading-off contacts' requests against each other. From this vantage point, an actor's social capital is seen as a function of the brokerage opportunities arising from the structural holes the actor's social relationships span [36, 109, 175].

Most research efforts in the area of online and offline social capital have focused on establishing the role of the Internet and social media in the accumulation of social capital [190, 67, 68]. Notably, the authors in [197] find that the usage of the Internet supplements social capital but does not increase or decrease it. More recent findings show that the use of social networking sites such as Facebook may in fact increase offline social capital by converting latent ties into online weak ties [67] and by allowing for a larger number of online heterogenous weak ties [174]. Users of Facebook and other social media have been associated with an increased social capital when compared to non-users [100, 174]. Our work builds on these findings, and further explores the structural properties of brokerage in two parallel online and offline networks. To the best of our knowledge it is the first work to study structural holes and brokerage in a multilayer context.

Geography plays an important role in the diversity of social networks where individuals with more geographically and structurally diverse networks are found to have higher social capital and come from more well-off areas in the UK [63]. The competitive advantage of an individual in a social network has been defined as a function of the structural holes that provide the individual with a brokerage position between otherwise disconnected others. [38]. Network brokerage has been studied in a variety of contexts from organisations [39] to online social networks [113], and most recently in geo-social networks [88]. The diversity of locations with respect to visitors has also been explored in the context of location-based social networks [50, 164]. We combine these approaches to consider the social network diversity of places and to the best of our knowledge, this is the first time that brokerage has been applied to interconnected networks [33, 22].

Urban Mobility & Deprivation. Human mobility in urban environments has received much attention in recent years, enabled in part by the increasing availability of individual-level data via online location-sharing services and mobile phone records. Human movement in urban areas differs from other geographic regions in their strong dependence on the spatial distribution of places within the city [135]. Location-sharing services have also been used to explore urban mobility through its impact on the place network [138]. While such studies focus on place, other work has explored the interplay between space and social ties, which are known to show strong inter-dependence [180]. In the context

of location-based social networks, a variety of geo-social phenomena have been analysed, including tie formation [9], co-location patterns [29, 44], homophily [201], and community structure [31]. The relationship between human mobility and the urban social diversity of places, however, has not been extensively studied before.

At present, government census studies are the most widely used measures of urban socioeconomic wellbeing at the neighbourhood level. However, recent works proposing the use of user-generated content and social media for measurement have emerged [111, 188, 151, 184] as part of a new science of cities, based on the availability of these new forms of data [16]. Using interim real-time measures of the urban pulse is appealing from a temporal and cost perspective but can be challenging due to the demographic biases of digital media users [61]. It is precisely these biases, however, that could provide insight into some of the most difficult to quantify and predict processes such as gentrification, which is associated with the displacement of residents of a deprived area by an influx of an (economically and digitally) affluent population. The link between urban social diversity and deprivation as measured by social media has yet not been studied.

Previous work using curated national statistical data sources has shown that the morphology of urban environments plays a key role in urban deprivation [185, 114], and that socioeconomic prosperity can be linked with other neighbourhood-level features such as human travel patterns [173], access to local facilities [114], and the prevalence of fast-food outlets [145]. Identifying and understanding associations such as these is of great interest to national policymakers, social reformers, and city planners. More recently, there has been interest in using signals from technological systems to predict urban wellbeing and deprivation. Many signals have proved useful in predicting deprivation indices, from passenger transits recorded by automated fare collection systems [168, 101] to crowdsourced data such as OpenStreetMap [188]. In the context of location-based networks, deprivation has been studied in terms of Foursquare's crowdsourced venue database [188, 152], but there has so far been little analysis from a joint geo-social perspective.

5.4 Discussion and Implications

The sociological tradition that places emphasis on gaps in social structures can be traced back to the late 1960s and early 1970s, when a group of sociologists began to develop the general idea that it is advantageous to forge connections to multiple, otherwise disconnected, individuals or groups [45, 72, 198]. One of the most celebrated theoretical endeavours in the social sciences that draws on this tradition is Granovetter's influential study of the bridging role of weak ties [79]. The broader the access to weaker ties, the closer an actor is likely to be placed to discontinuities in the social structure, which in turn enables the actor to be connected to various social circles of contacts and to be exposed to novel and diverse sources of information.

Brokerage has been studied in various contexts within the boundaries of a single network - e.g., organisational relations [40], online social network friendship [113], or mobile communication [63]. Brokerage, however, is not limited to one type of context. The same individuals may be engaged in different types of social relationships, and as a result may benefit from various brokerage positions that affect each other in complex ways. For instance, people can be brokers both online and offline. Recent studies have suggested that online social networking is directly linked to bridging and bonding social capital [68, 190], where social media sites such as Facebook have been shown to have a significant role in maintaining distant and near-by contacts, and in sustaining social capital [35]. However, while an individual may seem to be embedded in a socially cohesive neighbourhood online (offline), there may be opportunities for brokerage arising from the non-redundant contacts the individual has offline (online). *Despite the growing availability and ubiquity of online social media, it is still unclear whether there is a trade-off between the brokerage positions that individuals occupy online and offline.*

In this chapter, a multi-relational perspective on brokerage using a multilayer network approach to social capital was presented, where both online social network structure and physical co-location are taken into account to detect brokerage positions. With the goal of informing the advancement of location-based services and urban development in terms of deprivation and gentrification indices, in this chapter we also measure the bridging and bonding potential of various types of places and their visitor diversity using an interconnected network model of people and places. As opposed to a classic network model, this approach allows for social network projections of the spatial network and spatial network projections of the social network, enabling the measurement of the social properties of places and the geographical properties of people based on their place (spatial) network of visits. We use geo-social online data in Europe's largest metropolitan city – London – and introduce an interconnected people-place network paradigm, which aims to more realistically model urban social diversity. Using the Twitter social network of visitors to define the diversity of Foursquare venues, we are able to distinguish between categories, geographies and socioeconomic factors across London's neighbourhoods.

Venue recommendations have become a large part of urban discovery, especially for newcomers to the city. While aspects such as popularity or rating of a place are easily accessible through location-based services, the social role a place plays within the urban context is normally only well known by locals through experience. Whether a place is touristy or quiet, artsy or mainstream can be integrated into mobile system design for empowering newcomers or visitors to feel like locals. The role of serendipity and brokerage as described in the chapter can be particularly impactful when applied to location-based dating applications, where recommendations of places with new and diverse people play a fundamental role. Situation and mood dependent queries such as 'I want to have *drinks alone* at a place where it is *socially acceptable* and I can *meet new people*' will become increasingly popular given the social challenges cities present, in particular for the multi-

tude of newcomers. Despite the challenges of integrating data from different sources and running such metrics in real-time, the conceptual framework and urban social diversity metrics presented in this thesis can greatly benefit local businesses as well as innovative location-based discovery applications.

Another important implication of this work is its ability to underpin novel analysis of urban dynamics. The distribution of deprivation across neighbourhoods in relation to diversity is a topic, which affects local governments and policymakers. In particular, the finding that more socially cohesive and homogeneous communities tend to be either very wealthy or very poor but neighbourhoods with both high entropy and deprivation are the ones which are currently undergoing processes of gentrification is in agreement with previous literature on homophily where tightly knit communities are more resistant to change and resources remain within the community [123]. This suggests that affluent communities remain affluent and poor communities remain poor through isolation. On the other hand, areas where there is high diversity and deprivation, communities are undergoing change and what can be described as a gentrification process. This is confirmed by the sharp improvement of their IMD scores between 2010 and 2015. Although further investigation is needed into confounding factors and generalisability to other areas of the UK and the world, the inherent biases of social media demographics [61] work to the advantage of identifying a sudden rise in the affluence of visitors to a particularly deprived neighbourhood. Diversity metrics applied to social media data as in the present analysis can act as good predictors of gentrification when measured through indices of deprivation. Building applications that not only serve users and businesses but are also conscious of their impact on urban life in the longer run can become detrimental to urban development.

Chapter 6

Reflections and Outlook

Decades of disciplinary studies have advanced our understanding of urban and social dynamics where the preconception that only domain experts can tackle certain domain-specific problems has prevailed in science. However, with the emergence of new sources of geographic and social data, empirically grounded and large-scale studies of old and new problems have become possible through a combination of computational data-driven analysis and domain-specific theory. This thesis has demonstrated the use of such methods and more importantly has demonstrated some of the advantages of a multilayer approach to social and geographical systems.

The work presented in this dissertation is the product of two recent advancements: the increased availability of large-scale geo-social data and the rekindled interest in the modelling of multilayer networks. Leveraging both of these, a multilayer approach to the analysis of social, urban and international geo-social interactions was demonstrated, offering novel insights into these systems. By augmenting social networks with geographical interactions and place networks with social interaction data, this dissertation has applied the extension of traditional measures of centrality, diversity and overlap as well as new perspectives on geo-social networks not possible without the large-scale contribution of online data streams.

The approach taken has been incremental, first introducing a minimal multiplex network model for measuring global connectivity and socioeconomic similarity using a number of physical and digital networks, followed by a multiplex model of tie strength in small but interaction-rich social multiplexes and larger-scale geo-social networks. The aim has been to fill the gap in multilayer literature between theoretical and empirical research and to present a data-driven approach which reflects the current state of geo-social network data as produced by online and offline systems. Through this perspective, new models of multilayer interactions are also needed and an interconnected geo-social network was presented to this end, which demonstrates how people-place interactions can be modelled to gain novel insights about the dynamics of cities.

6.1 Summary of Contributions

The thesis that *a multilayer approach to urban and social theory can advance our understanding of geo-social networks beyond what is possible from studying their social and geographical components in isolation* has been substantiated through the following contributions presented in this dissertation:

Approximating socioeconomic indicators through multiplex interactions: In Chapter 3, we introduced a simple multiplex model for the analysis of complex systems applied to international relations. We built a global multiplex as a collection of graphs of the postal, trade, migration, flights, digital communication and physical Internet layers. These physical and digital layers of international resource flows were combined and studied together through the multiplex framework, revealing novel insight about the distribution of wealth and other resources and showing that a combined multiplex degree correlates better with critical socio-economic indicators. We also introduced the *community multiplexity index*, a measure of community membership similarity, and applied it to approximating the socioeconomic profiles of countries, which allows for critical indicators for international development purposes to be estimated from community similarity when such data is missing for a particular country.

Homophily in multirelational networks: Influential sociologists have long called for a dimensional exploration of homophily, both in terms of similarity characteristics and methodology. In Chapter 4, we applied the concept of multiplexity to the study of homophily within a student community. Using multiplex weighted distance, measuring the multichannel communication of students and the strength of their ties, we showed that those at closer distance in the multiplex network, were more similar than those further away, evaluating a number of dimensions such as music taste, political affiliation, situational factors and health habits. This is also a validation of the *media multiplexity* theory, which has not been examined in the multilayer context or applied as a measure of multiplex tie strength.

Link prediction in multilayer online networks: In Chapter 4, we further explored the classic problem of link prediction in social networks from the multilayer perspective. We formulated a number of location-based, social and mixed features which we use to predict Twitter links from Foursquare features and vice versa. Additionally, in a second task we demonstrate how mixed multilayer features can be used to predict which links exist across layers, extending current link prediction literature to heterogeneous cross-platform prediction and introducing the problem of multilayer link prediction for the first time.

Social capital and structural holes in the multilayer context: The concept of social capital is explored in Chapter 5 from a geo-social perspective, where the notion that social capital can be generated on both offline geographic and online social networks is presented. It is then empirically observed that by considering just one layer - social or geographic - social capital can be under or over estimated. A unified view of brokerage for measuring social capital across online and offline layers of interaction was presented to ameliorate this problem. This is the first study of social capital, brokerage and structural holes from a multilayer perspective.

An interconnected network of people and places: With the goal of providing realistic modelling for cities as they face more and more complex problems due to overpopulation and globalisation, Chapter 5 introduces an interconnected model of geo-social interactions between people and places. This model combines the social network of citizens with the place network of places around the city. The concept of brokerage is then redefined for places in terms of redundancy in the social network of its visitors. We also introduce the concept of serendipity as the probability of the social composition of a place, given the previously visited place network of each visitor. Ultimately, these measures of urban diversity were compared to indices of deprivation for London, where neighbourhoods which are characterised by high social diversity in terms of the interconnected network model but have high deprivation, were identified as neighbourhoods which are undergoing processes of change such as gentrification. This was validated by observing the percent change in terms of improvement of these neighbourhoods as well as independent urban research pointing to those neighbourhoods as gentrifying.

6.2 Future Directions & Outlook

This dissertation has presented a data-driven computational approach to multilayer networks, which has been under-explored in the predominantly theoretical field of multilayer complex networks. The domain of geo-social networks is a particularly fascinating one as it holds the key to understanding the online and offline human dynamics that govern our modern world. With the new availability of geographic and social interaction data, studying how humans navigate the virtual and physical world in parallel will become increasingly popular. Moreover, interfaces which intersect both dimensions will necessitate the use of models beyond the single-layer, shifting data-driven multilayer networks from an approach to a necessary method of evaluating multidimensional data.

The multilayer approach demonstrated in extending the theories of homophily, media multiplexity and brokerage in this dissertation is not limited to these theories and can be widely applied in cross-disciplinary research such as computational social science. Theories such as Bourdieu's cultural capital [24] can be applied as extensions of Chapter 5,

where not only social capital but also the economic and cultural capital of individuals and places can be studied in parallel. Further candidates for multilayer analysis are Hofstede's cultural dimensions [87], which can be seen as a layer of the network of cross-cultural interaction each and where the pattern of each culture can be characterised as a multilayer construct, as well as Campbell's hero's journey patterns of storytelling [42] where an interconnected network can be used to describe the causality and cascading effects in applications to large-scale language processing. All of these examples have in common multiple factors which contribute to their formation but this dimensionality has not been studied as a network yet. This dissertation has not explored time and dimension varying multilayer networks which can prove useful constructs in understanding these theories in a data-driven computational manner. Revisiting our understanding of social and geographical systems from a multilayer perspective, as this body of research has hopefully shown, can provide a deeper perspective on the broader human condition.

Future directions in multilayer analysis will present themselves in many other domains where massive amounts of data are collected such as ecology, organisational studies and marketing. With one of the biggest challenges we face being climate change, more efforts of collecting data about CO_2 emissions and pollution are made. Studying these from a geographic and human activity perspective as a multilayer interconnected network can shed light on the processes driving pollution and key measures to prevent it. Furthermore, organisational studies have long explored hierarchies of companies but as organisations become more flat, new models for understanding how groups form in an ad-hoc fashion in open office spaces are needed. Multilayer perspectives on groups and more in general on human activity can increase the effectiveness of marketing strategies as people and places redefine each other as has been suggested in Chapter 5 of this dissertation.

From an application-based perspective, as discussed in Chapter 4, there is more and more integration between online services which provides rich user interaction data on many levels. Apart from the obvious marketing value of this information, it will become increasingly important to have seamless integration between devices and services, which poses a number of security threats as well as design opportunities. Identifying users across online social media platforms is still a developing research thread [77], which will also manifest itself in better recommendations based on multidimensional data from across services. For example, obtaining a person's interests from Pinterest, their social network from Facebook and mobility network from Foursquare could lead to a revolution in personalised non-transient recommendations such as where to live or work in order to augment your social and cultural capital as opposed to short-term item based recommendations.

The idealised concept of a "smart city" [16], where technological advancements allow for fluid and seamless interactions between citizens and cities is the focus of architects, designers, governments and computer scientists alike. However, in order to build the cities of the future together, common conceptual models are needed to advance ideas into

practice. In Chapter 5 we introduced such a model, which is generalisable to different types of people-place interactions. This can be extended, for example, to understand optimal urban transit strategies by adding an intermediary layer of transportation as a medium of transition between people and places. Ultimately, the most important aspect of a smart city is that it is adaptable to its resident's ever-changing needs and interconnected network models could be the common language and insight which relates to these needs.

All together these chapters suggest that multilayer modelling of geo-social data is an effective way to gain novel insights about the world and has put forward a framework which aims to inspire more interdisciplinary research and data-driven multilayer research not only in the realm of geo-social networks but also beyond.

Bibliography

[1] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. The network origins of large economic downturns. Technical report, National Bureau of Economic Research, 2013.

[2] Lada Adamic and Eytan Adar. Friends and neighbors on the Web. *Social Networks*, 25:211–230, 2001.

[3] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6(2):9:1–9:33, 2012.

[4] José Anson and Matthias Helblei. *A gravity model of international postal exchanges*. Edward Elgar Publishing, 2013.

[5] Jan Ketil Arnulf, Kai Rune Larsen, Øyvind Lund Martinsen, and Chih How Bong. Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PloS ONE*, 9(9):e106361, 2014.

[6] Lars Backstrom, Eytan Bakshy, Jon Kleinberg, Thomas Lento, and Itamar Rosen. Center of attention: How Facebook users allocate attention across friends. In *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM'11)*, 2011.

[7] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42. ACM, 2012.

[8] Lars Backstrom and Jure Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *WSDM*. ACM, 2011.

[9] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 61–70, 2010.

[10] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st International*

Conference on World Wide Web (WWW'12), pages 519–528, New York, NY, USA, 2012. ACM.

[11] Albert-László Barabási. The network takeover. *Nature Physics*, 8:14–16, 2013.

[12] Matteo Barigozzi, Giorgio Fagiolo, and Diego Garlaschelli. Multinetwork of international trade: A commodity-specific analysis. *Physical Review E*, 81(4):046104, 2010.

[13] Matteo Barigozzi, Giorgio Fagiolo, and Giuseppe Mangioni. Community structure in the multi-network of international trade. In *Complex Networks*, pages 163–175. Springer, 2011.

[14] Louise Barrett, S. Peter Henzi, and David Lusseau. Taking sociality seriously: the structure of multi-dimensional social networks as a source of information for individuals. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1599):2108–2118, 2012.

[15] Federico Battiston, Vincenzo Nicosia, and Vito Latora. Structural measures for multiplex networks. *Physical Review E*, 89(3):032804, 2014.

[16] Michael Batty. *The New Science of Cities*. MIT Press, 2013.

[17] Michael Batty and Yichun Xie. From cells to cities. *Environment and Planning B*, 21:s31–s31, 1994.

[18] Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Multidimensional networks: foundations of structural analysis. *World Wide Web*, 16(5-6):567–593, 2013.

[19] Ginestra Bianconi. Statistical mechanics of multiplex networks: Entropy and overlap. *Physical Review E*, 87, 2013.

[20] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 10, 2008.

[21] Jeffrey Boase and Barry Wellman. Personal relationships: On and off the Internet. In *The Cambridge Handbook of Personal Relationships*, pages 709–724. Cambridge University Press, 2006.

[22] Stefano Boccaletti, G Bianconi, R Criado, Charo I Del Genio, J Gómez-Gardeñes, M Romance, I Sendina-Nadal, Z Wang, and M Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[23] Stephen P. Borgatti. Structural holes: Unpacking burt's redundancy measures. *Connections*, 20:35–38, 1997.

[24] Pierre Bourdieu. The forms of capital (1986). *Cultural theory: An anthology*, pages 81–93, 2011.

[25] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[26] Piotr Bródka and Przemyslaw Kazienko. Multi-layered social networks. *CoRR*, abs/1212.2425, 2012.

[27] Piotr Bródka, Katarzyna Musial, and Przemysław Kazienko. A method for group extraction in complex social networks. In *World Summit on Knowledge Society*, pages 238–247. Springer, 2010.

[28] Piotr Bródka, Krzysztof Skibicki, Przemysław Kazienko, and Katarzyna Musiał. A degree centrality in multi-layered social network. In *International Conference on Computational Aspects of Social Networks (CASoN)*, pages 237–242. IEEE, 2011.

[29] Chloë Brown, Neal Lathia, Cecilia Mascolo, Anastasios Noulas, and Vincent Blondel. Group colocation behavior in technological social networks. *PLoS ONE*, 9(8), August 2014.

[30] Chloë Brown, Vincenzo Nicosia, Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. The importance of being placefriends: discovering location-focused online communities. In *Proceedings of the ACM workshop on Online Social Networks*, pages 31–36. ACM, 2012.

[31] Chloë Brown, Vincenzo Nicosia, Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Social and place-focused communities in location-based online social networks. *The European Physical Journal B*, 86(6):1–10, June 2013.

[32] Charles D. Brummitt, Raissa M. DSouza, and E. A. Leicht. Suppressing cascades of load in interdependent networks. *Proceedings of the National Academy of Sciences*, 2012.

[33] Sergey V Buldyrev, Roni Parshani, Gerald Paul, H Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010.

[34] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[35] Moira Burke and Robert Kraut. Using Facebook after losing a job: Differential benefits of strong and weak ties. *Proceedings of CSCW'13*, 2013.

[36] R. S. Burt. *Structural Holes: The Social Structure of Competition.* Harvard University Press, Cambridge MA, 1992.

[37] Ronald S Burt. *Structural Holes*. Harvard University Press, 1995.

[38] Ronald S Burt. The network structure of social capital. *Research in Organizational Behavior*, 22:345–423, 2000.

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

[40] Ronald S Burt. *Brokerage and Closure: An Introduction to Social Capital*. Oxford University Press, New York and Oxford, 2005.

[41] Frances Cairncross. *The death of distance: How the communications revolution is changing our lives*. Harvard Business Press, 2001.

[42] J. Campbell. *The Hero with a Thousand Faces*. Bollingen series. New World Library, 2008.

[43] Alessio Cardillo, Massimiliano Zanin, Jesus Gomez-Gardenes, Miguel Romance, Alejandro J. Garcia del Amo, and Stefano Boccaletti. Modeling the multi-layer nature of the European air transport network: Resilience and passengers re-scheduling under random failures. *The European Physical Journal Special Topics*, 215(1):23–33, 2013.

[44] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining*, pages 1082–1090. ACM, 2011.

[45] K.S. Cook and R.M. Emerson. Power, equity and commitment in exchange networks. *American Sociological Review*, 43:712–739, 1978.

[46] Emanuele Cozzo, Mikko Kivelä, Manlio De Domenico, Albert Solé-Ribalta, Alex Arenas, Sergio Gómez, Mason A Porter, and Yamir Moreno. Structure of triadic relations in multiplex networks. *New Journal of Physics*, 17(7):073029, 2015.

[47] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. In *Proceedings of the National Academy of Sciences*, volume 107, pages 22436–22441, 2010.

[48] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining*, pages 160–168. ACM, 2008.

[49] Justin Cranshaw, Raz Schwartz, Jason I Hong, and Norman Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *International AAAI Conference on Weblogs and Social Media*, page 58, 2012.

[50] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Ubicomp*, pages 119–128, 2010.

[51] Paul Craven and Barry Wellman. The network city*. *Sociological inquiry*, 43(3-4):57–88, 1973.

[52] Manlio De Domenico, Andrea Lancichinetti, Alex Arenas, and Martin Rosvall. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Physical Review X*, 5(1):011027, 2015.

[53] Manlio De Domenico, Antonio Lima, Paul Mougel, and Mirco Musolesi. The anatomy of a scientific rumor. *Scientific reports*, 3, 2013.

[54] Manlio De Domenico, Vincenzo Nicosia, Alexandre Arenas, and Vito Latora. Structural reducibility of multilayer networks. *Nature communications*, 6, 2015.

[55] Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A Porter, Sergio Gómez, and Alex Arenas. Mathematical formulation of multilayer networks. *Physical Review X*, 3, 2013.

[56] Manlio De Domenico, Albert Solé-Ribalta, Sergio Gómez, and Alex Arenas. Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences*, 111(23):8351–8356, 2014.

[57] Manlio De Domenico, Albert Solé-Ribalta, Elisa Omodei, Sergio Gómez, and Alex Arenas. Ranking in interconnected multilayer networks reveals versatile nodes. *Nature communications*, 6, 2015.

[58] Pierre Deville, Chaoming Song, Nathan Eagle, Vincent D Blondel, Albert-László Barabási, and Dashun Wang. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences*, 2016.

[59] W. Dong, B. Lepri, and A. Pentland. Modelling the co-evolution of behaviours and social relationships using mobile phone data. In *Proceedings of MUM'11*, pages 134–143, 2011.

[60] Jonathan F Donges, Hanna CH Schultz, Norbert Marwan, Yong Zou, and Jürgen Kurths. Investigating the topology of interacting networks. *The European Physical Journal B*, 84(4):635–651, 2011.

[61] Maeve Duggan, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. Social media update 2014. *Pew Research Center*, 2015.

[62] Robin Dunbar. How many friends can you really have? *IEEE Spectrum*, pages 81–83, 2011.

[63] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[64] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[65] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[66] Steeve Ebener, Christopher Murray, Ajay Tandon, and Christopher C Elvidge. From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery. *International Journal of Health Geographics*, 4(1):1, 2005.

[67] N. B. Ellison, C. Steinfield, and C. Lampe. Connection strategies: Social capital implications of Facebook-enabled communication practices. *New Media & Society*, 20(10):1–20, 2011.

[68] N. B. Ellison, J. Vitak, R. Gray, and C. Lampe. Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *Journal of Computer-Mediated Communication*, 19(4):855–870, 2014.

[69] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[70] Claude S Fischer. *To dwell among friends : personal networks in town and city.* Chicago : University of Chicago Press, 1982.

[71] Richard L. Florida. *The Rise of the Creative Class: Revisited.* Basic Books, 2012.

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

[73] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Softw. Pract. Exper.*, 21(11):1129–1164, November 1991.

[74] Riccardo Gallotti and Marc Barthelemy. The multilayer temporal network of public transport in great britain. *Scientific data*, 2, 2015.

[75] David Garcia, Pavlin Mavrodiev, and Frank Schweitzer. Social resilience in online communities: The autopsy of friendster. In *Proceedings of the first ACM conference on Online social networks*, pages 39–50. ACM, 2013.

[76] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *CHI*, 2009.

[77] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*, pages 447–458, New York, NY, USA, 2013. ACM.

[78] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[79] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[80] Sébastian Grauwin, Stanislav Sobolevsky, Simon Moritz, István Gódor, and Carlo Ratti. Towards a comparative science of cities: Using mobile traffic records in New York, London, and Hong Kong. In *Computational approaches for urban environments*, pages 363–387. Springer, 2015.

[81] R Guimera, S Mossa, A Turtschi, and LAN Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities global roles. *Proceedings of the National Academy of Sciences*, 102(22), 2005.

[82] Arda Halu, Satyam Mukherjee, and Ginestra Bianconi. Emergence of overlap in ensembles of spatial multiplexes and statistical mechanics of spatial interacting network ensembles. *Physical Review E*, 89(1):012806, 2014.

[83] Caroline Haythornthwaite. Social Networks and Internet Connectivity Effects. *Information, Communication & Society*, 8(2):125–147, 2005.

[84] Caroline Haythornthwaite and Barry Wellman. Work, friendship, and media use for information exchange in a networked organization. *Journal of the American Society for Information Science*, 49(12):1101–1114, 1998.

[85] César A Hidalgo, Bailey Klinger, A-L Barabási, and Ricardo Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

[86] Russell A Hill and Robin Dunbar. Social network size in humans. *Human nature*, 14(1):53–72, 2003.

[87] Geert Hofstede. Cultural constraints in management theories. *The Academy of Management Executive*, 7(1):81–94, 1993.

[88] Desislava Hristova, Pietro Panzarasa, and Cecilia Mascolo. Multilayer brokerage in geo-social networks. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM'15)*, 2015.

[89] S.P. Huntington. *The Clash of Civilizations and the Remaking of World Order*. A Touchstone book. Simon & Schuster, 1996.

[90] Jacopo Iacovacci, Zhihao Wu, and Ginestra Bianconi. Mesoscopic structures reveal the network between the layers of multiplex data sets. *Physical Review E*, 92(4):042806, 2015.

[91] Danesh Irani, Steve Webb, Kang Li, and Calton Pu. Large online social footprints–an emerging threat. In *International Conference on Computational Science and Engineering*, volume 3, pages 271–276. IEEE, 2009.

[92] Jane Jacobs. *The Death and Life of Great American Cities*. Vintage International. Random House, 1961.

[93] Pablo Kaluza, Andrea Kölzsch, Michael T. Gastner, and Bernd Blasius. The complex network of global cargo ship movements. *Journal of The Royal Society Interface*, 7(48):1093–1103, 2010.

[94] Bruce Kapferer. *Strategy and transaction in an African factory: African workers and Indian management in a Zambian town*. Manchester University Press, 1972.

[95] P. Kazienko, P. Brodka, K. Musial, and J. Gaworecki. Multi-layered social network creation based on bibliographic data. In *SocialCom*, Aug 2010.

[96] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[97] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[98] Vassilis Kostakos, Eamonn O'Neill, Alan Penn, George Roussos, and Dikaios Papadongonas. Brief encounters: Sensing, modeling and visualizing urban mobility and copresence networks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(1):2, 2010.

[99] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 591–600, 2010.

[100] Cliff Lampe, Jessica Vitak, and Nicole Ellison. Users and nonusers: Interactions between levels of adoption and social capital. In *Proceedings of CSCW'13*, pages 809–820, New York, NY, USA, 2013. ACM.

[101] Neal Lathia, Daniele Quercia, and Jon Crowcroft. The hidden image of the city: Sensing community well-being from urban mobility. In *Pervasive Computing*, number 7319 in Lecture Notes in Computer Science, pages 91–98. Springer Berlin Heidelberg, June 2012.

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

[103] Kisung Lee, Raghu K. Ganti, Mudhakar Srivatsa, and Ling Liu. When Twitter meets Foursquare: Tweet location prediction using Foursquare. In *UbiComp*, 2014.

[104] Loretta Lees. Gentrification and social mixing: Towards an inclusive urban renaissance? *Urban Studies*, 45(12):2449–2470, 2008.

[105] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using Facebook. com. *Social networks*, 30(4):330–342, 2008.

[106] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[107] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[108] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining*, pages 243–252. ACM, 2010.

[109] N. Lin. *Social capital: A theory of social structure and action*. Cambridge University Press, New York, 2001.

[110] Richard G Little. Controlling cascading failure: understanding the vulnerabilities of interconnected infrastructures. *Journal of Urban Technology*, 9(1):109–123, 2002.

[111] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PLoS ONE*, 10, 05 2015.

[112] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PloS ONE*, 10(5):e0128692, 2015.

[113] Tiancheng Lou and Jie Tang. Mining structural hole spanners through information diffusion in social networks. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*, pages 825–836, 2013.

[114] Sally Macintyre, Laura Macdonald, and Anne Ellaway. Do poorer people have poorer access to local resources and facilities? The distribution of local resources by area deprivation in Glasgow, Scotland. *Social Science & Medicine*, 67(6):900–914, September 2008.

[115] George MacKerron and Susana Mourato. Happiness is greater in natural environments. *Global Environmental Change*, 23(5):992–1000, 2013.

[116] Sofus A. Macskassy and Matthew Michelson. Why do people retweet? Anti-homophily wins the day! In *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM'11)*, 2011.

[117] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. Pentland. Sensing the 'Health State' of a Community. *Pervasive Computing*, 11(4):36–45, 2012.

[118] A. Madan, K. Farrahi, D. Gatica-Perez, and A. Pentland. Pervasive sensing to model political opinions in face-to-face networks. In *Proceedings of Pervasive'11*, pages 214–231, 2011.

[119] Matteo Magnani and Luca Rossi. Formation of multiple networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 257–264. Springer, 2013.

[120] Gabriel Magno and Ingmar Weber. International gender differences and gaps in online social networks. In *Social Informatics*, pages 121–138. Springer, 2014.

[121] Huina Mao, Xin Shuai, Yong-Yeol Ahn, and Johan Bollen. Mobile communications reveal the regional economy in Côte dIvoire. *Proceedings of NetMob*, 2013.

[122] Noah Mark. Birds of a feather sing together. *Social Forces*, 77(2):pp. 453–485, 1998.

[123] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:215–444, 2001.

[124] Giulia Menichetti, Daniel Remondini, and Ginestra Bianconi. Correlations between weights and overlap in ensembles of weighted multiplex networks. *Physical Review E*, 90(6):062817, 2014.

[125] Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases*, pages 437–452. Springer, 2011.

[126] Byungjoon Min, Su Do Yi, Kyu-Min Lee, and K-I Goh. Network robustness of multiplex networks with interlayer degree correlations. *Physical Review E*, 89(4):042811, 2014.

[127] Diana Mok, Barry Wellman, and Juan Carrasco. Does distance matter in the age of the Internet? *Urban Studies*, 47(13):2747–2783, 2010.

[128] J.L. Moreno and H.H. Jennings. *Who shall survive?: A new approach to the problem of human interrelations*. Nervous and mental disease monograph series. Nervous and mental disease publishing co., 1934.

[129] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[130] United Nations. A world that counts: Mobilising the data revolution for sustainable development. White paper, 2014.

[131] M. E. J. Newman. Mixing Patterns in Networks. *Phys. Rev. E*, 67:026126, Feb 2003.

[132] Mark Newman, Albert-Laszlo Barabasi, and Duncan J. Watts. *The Structure and Dynamics of Networks: Princeton Studies in Complexity*. Princeton University Press, Princeton, NJ, USA, 2006.

[133] Vincenzo Nicosia, Ginestra Bianconi, Vito Latora, and Marc Barthelemy. Growing multiplex networks. *Physical review letters*, 111(5):058701, 2013.

[134] Vincenzo Nicosia and Vito Latora. Measuring and modeling correlations in multiplex networks. *Physical Review E*, 92(3):032805, 2015.

[135] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE*, 7(5):e37027, May 2012.

[136] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, Security, Risk and Trust (PASSAT), International Confernece on Social Computing (SocialCom)*, pages 144–153. IEEE, 2012.

[137] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM'11)*, 11:70–573, 2011.

[138] Anastasios Noulas, Blake Shaw, Renaud Lambiotte, and Cecilia Mascolo. Topological properties and temporal dynamics of place networks in urban environments. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*, pages 431–441, 2015.

[139] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332, 2007.

[140] Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. Geographic constraints on social network groups. *PLoS ONE*, 6(4):e16939, 2011.

[141] Raphael Ottoni, Diego Las Casas, Joo Paulo Pesce, Wagner Meira Jr., Christo Wilson, Alan Mislove, and Virgilio Almeida. Of pins and tweets: Investigating how users behave across image- and text-based social networks. In *Proceedings of the 8th International AAAI Conference on Web and Social Media (ICWSM'14)*, 2014.

[142] John F. Padgett and Paul D. Mclean. Organizational invention and elite transformation: The birth of partnership systems in Renaissance Florence. *American Journal of Sociology*, 111:1463–1568, 2006.

[143] F. Papadopoulos, M. Kitsak, M. Serrano, M. Bogu, and D. Krioukov. Popularity Versus Similarity in Growing Networks. *Nature*, 489:537–540, Sept 2012.

[144] Luca Pappalardo, Giulio Rossetti, and Dino Pedreschi. "how well do we know each other?" Detecting tie strength in multidimensional social networks. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1040–1045. IEEE, 2012.

[145] Jamie Pearce, Tony Blakely, Karen Witten, and Phil Bartie. Neighborhood Deprivation and Access to Fast-Food Retailing: A National Study. *American Journal of Preventive Medicine*, 32(5):375–382, May 2007.

[146] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[147] Tiago P Peixoto. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807, 2015.

[148] Plato. *Plato in Twelve Volumes*. Number v. 11 in Loeb Classical Library. Harvard University Press, 1968.

[149] United Nations Global Pulse. Big data for development: Challenges & opportunities. White paper, 2012.

[150] Robert D. Putnam. *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, New York, 2000.

[151] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. Tracking gross community happiness from tweets. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW'12)*, pages 965–968. ACM, 2012.

[152] Daniele Quercia and Diego Saez. Mining urban deprivation from Foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing*, 13(2):30–36, April 2014.

[153] Daniele Quercia, Diarmuid O Seaghdha, and Jon Crowcroft. Talk of the city: Our tweets, our community happiness. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[154] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67, 2003.

[155] Jean-Paul Rodrigue, Claude Comtois, and Brian Slack. *The geography of transport systems*. Routledge, 2013.

[156] Giulio Rossetti, Michele Berlingerio, and Fosca Giannotti. Scalable link prediction on multidimensional networks. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 979–986. IEEE, 2011.

[157] Alex Rutherford, Manuel Cebrian, Iyad Rahwan, Sohan Dsouza, James McInerney, Victor Naroditskiy, Matteo Venanzi, Nicholas R Jennings, Eero Wahlstedt, Steven U Miller, et al. Targeted social mobilization in a global manhunt. *PloS one*, 8(9):e74628, 2013.

[158] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. Finding your friends and following them to where you are. In *WSDM*, 2012.

[159] S.F. Sampson. *A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships*. 1968.

[160] Rubén J Sánchez-García, Emanuele Cozzo, and Yamir Moreno. Dimensionality reduction and spectral properties of multilayer networks. *Physical Review E*, 89(5):052815, 2014.

[161] Anna Saumell-Mendiola, M Ángeles Serrano, and Marián Boguñá. Epidemic spreading on interconnected networks. *Physical Review E*, 86(2):026106, 2012.

[162] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Jon Crowcroft. Track globally, deliver locally: Improving content delivery networks by tracking geographic social cascades. In *Proceedings of the 20th International Conference on World Wide Web*, pages 457–466. ACM, 2011.

[163] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM'11)*, 11:329–336, 2011.

[164] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining*, pages 1046–1054, 2011.

[165] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining*, 2011.

[166] Maximilian Schich, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. A network framework of cultural history. *Science*, 345(6196):558–562, 2014.

[167] Yuval Shavitt and Eran Shir. Dimes: Let the Internet measure itself. *ACM SIGCOMM Computer Communication Review*, 35(5):71–74, 2005.

[168] Chris Smith, Daniele Quercia, and Licia Capra. Finger on the pulse: Identifying deprivation using transit flow analysis. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW'13)*, CSCW '13, pages 683–692, 2013.

[169] Christopher Smith-Clarke, Afra Mashhadi, and Licia Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 511–520. ACM, 2014.

[170] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[171] Victor Soto, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 377–388. Springer, 2011.

[172] Bogdan State, Patrick Park, Ingmar Weber, Michael Macy, et al. The mesh of civilizations in the global network of digital communication. *PloS ONE*, 10(5):e0122543, 2015.

[173] Dominic Stead. Relationships between land use, socioeconomic factors, and travel patterns in Britain. *Environment and Planning B: Planning and Design*, 28(4):499–528, August 2001.

[174] Charles Steinfield, Joan M DiMicco, Nicole B Ellison, and Cliff Lampe. Bowling online: Social networking and social capital within the organization. In *Proceedings of the Fourth International Conference on Communities and Technologies*, pages 245–254. ACM, 2009.

[175] K. Stovel and L. Shaw. Brokerage. *Annual Review of Sociology*, 38:139–158, 2012.

[176] Alistair Sutcliffe, Robin Dunbar, Jens Binder, and Holly Arrow. Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology*, 103(2):149–168, 2012.

[177] Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010.

[178] Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties Across Heterogenous Networks. In *WSDM*, 2012.

[179] LB Hackney Policy Team. A profile of Hackney, its people and place. *Hackney Council*, 2014.

[180] Jameson L. Toole, Carlos Herrera-Yaqüe, Christian M. Schneider, and Marta C. González. Coupling human mobility and social ties. *Journal of The Royal Society Interface*, 12(105):20141128, April 2015.

[181] Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[182] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the Facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[183] Electronic Postal Services Program Universal Postal Union. Measuring postal e-services development, 2012.

[184] Carmen Vaca Ruiz, Daniele Quercia, Luca Maria Aiello, and Piero Fraternali. Taking Brazil's pulse: Tracking growing urban economies from online attention. In

Proceedings of the 23rd International Conference on World Wide Web (WWW'14), pages 451–456, 2014.

[185] Laura Vaughan, David L Chatford Clark, Ozlem Sahbaz, and Mordechai Haklay. Space and exclusion: Does urban morphology play a part in social deprivation? *Area*, 37(4):402–412, 2005.

[186] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the Internet. *Physical Review E*, 65, 2002.

[187] Alexei Vazquez. Spreading dynamics on heterogeneous populations: Multitype network approach. *Physical Review E*, 74(6):066114, 2006.

[188] Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. Measuring urban deprivation from user generated content. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'15)*, pages 254–264, 2015.

[189] Lois M. Verbrugge. Multiplexity in adult friendships. *Social Forces*, 57(4):1286–1309, 1979.

[190] J. Vitak, N. B. Ellison, and C. Steinfield. The ties that bond: Re-examining the relationship between Facebook use and bonding social capital. In *Proceedings of the 44th Hawaii International Conference on System Sciences*, pages 1–10, 2011.

[191] Jessica Vitak. The Impact of Context Collapse and Privacy on Social Network Site Disclosures. *Journal of Broadcasting & Electronic Media*, 56(4):451–470, 2012.

[192] Jessica Vitak. Facebook makes the heart grow fonder: Relationship maintenance strategies among geographically dispersed and communication-restricted connections. In *CSCW*, pages 842–853, 2014.

[193] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining*, 2011.

[194] Zhen Wang, Lin Wang, and Matjaž Perc. Degree mixing in multilayer networks impedes the evolution of cooperation. *Physical Review E*, 89(5):052813, 2014.

[195] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[196] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440, 1998.

[197] B. Wellman, A. Q. Haase, J. Witte, and K. Hampton. Does the Internet increase, decrease, or supplement social capital? social networks, participation, and community commitment. *American Behavioral Scientist*, 45(3):436–455, 2001.

[198] Harrison C White. *Chains of Opportunity*. Harvard University Press, Cambridge, MA, 1970.

[199] Yang Yang, Niran Chawla, Yizhou Sun, and Jiawei Hani. Predicting links in multi-relational and heterogeneous networks. In *12th International Conference on Data Mining (ICDM)*, pages 755–764. IEEE, 2012.

[200] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. Detecting spam in a twitter network. *First Monday*, 15(1), 2009.

[201] Ke Zhang and Konstantinos Pelechrinis. Understanding spatial homophily: The case of peer influence and social selection. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, pages 271–282, 2014.

[202] Changtao Zhong, Mostafa Salehi, Sunil Shah, Marius Cobzarenco, Nishanth Sastry, and Meeyoung Cha. Social bootstrapping: How Pinterest and Last.Fm social communities benefit by borrowing links from Facebook. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, 2014.