# COMPUTATIONAL LENS ON BIG SOCIAL AND INFORMATION NETWORKS

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Yuxiao Dong

---

Nitesh V. Chawla, Director

Graduate Program in Computer Science and Engineering
Notre Dame, Indiana

March 2017

COMPUTATIONAL LENS ON BIG SOCIAL AND INFORMATION NETWORKS

Abstract

by

Yuxiao Dong

The connections between individuals form the structural backbone of human societies, which manifest as networks. In a network sense, individuals matter in the ways in which their unique demographic attributes and diverse interactions activate the emergence of new phenomena at larger, societal levels. Accordingly, this thesis develops computational models to investigating the ways that individuals are embedded in and interact within a wide range of over one hundred big networks—the biggest with over 60 million nodes and 1.8 billion edges—with an emphasis on two fundamental and interconnected directions: user demographics and network diversity.

Work in this thesis in the direction of demographics unveils the social strategies that are used to satisfy human social needs evolve across the lifespan, examines how males and females build and maintain similar or dissimilar social circles, and reveals how classical social theories—such as weak/strong ties, social balance, and small worlds—are influenced in the context of digitally recorded big networks coupled with socio-demographics. Our work on demographics also develops scalable graphical models that are capable of incorporating structured discoveries (features), facilitating conventional data mining tasks in networks. Work in this part demonstrates the predictability of user demographic attributes from networked systems, enabling the potential for precision marketing and business intelligence in social networking services. Work in this thesis in the direction of diversity examines how the

diverse structures of common neighborhood influence link formation locally and network organization globally, how this influence varies across different types of social and information networks, and how it concords or conflicts with the principle of homophily. Work in this direction reveals how topic diversity—in contrast to authority and popularity—drives the growth of impact in academic collaboration and citation networks as well. Finally, our work on diversity presents neural network based representation learning models for embedding heterogeneous networks in which there exist diverse types of nodes and edges, giving rise to important implications for traditional mining and learning tasks in heterogeneous network data, including similarity search, clustering, and classification.

Dedicated to all networked beings.

CONTENTS

FIGURES

TABLES

# ACKNOWLEDGMENTS

First, I would like to thank my advisor, Professor Nitesh V. Chawla, for his tremendous help and support over the course of my Ph.D. studies. I am greatly indebted to him for the vision, encouragement, and freedom that he has generously provided. I can also never overstate my gratitude for his approval and support for many of my London travels and remote work, from which my relationship with Yeki has largely benefited. I would also like to thank David Chiang (CSE), Omar Lizardo (Sociology), and Zoltán Toroczkai (Physics) for serving on my thesis committee and providing deep and diverse comments. Omar's remarkably insightful thoughts and timely executions have inspired me in many ways. Working with him has always been instrumental to furthering my understanding of networks from a sociological perspective, and every discussion with him has always ended with me learning new things. Over my time in 384 Nieuwland Hall, I have observed Zoltán to be the perfect personification of the inquiring scientist, precisely as I envisioned during my youth.

This thesis could not be possible without my amazing group of collaborators. I am grateful for them: Francesco Calabrese, Xiaoming Fu, Hong Huang, Reid Johnson, Omar, Tiancheng Lou, Hao Ma, Qiaozhu Mei, Fabio Pinelli, Iris Shen, Ananthram Swami, Jie Tang, Bai Wang, Kuansan Wang, Sen Wu, Xian Wu, Jian Xu, Yang Yang, Yang Yang (ND), and Jing Zhang. I would like to give my unbounded thanks to Jie for his unbelievably generous mentorship and friendship. I am particularly fortunate to begin working with him before I came to the States, during which I was exposed to the exciting social networks research. I am more than grateful for the amount of time and insightful feedback he has consistently shared with me since then, as well

as our jogging competitions and shoppings in so many places: Beijing, New York, San Francisco, Redmond, Indianapolis, and Notre Dame. When looking back at this thesis and my life at Notre Dame, I feel that I am overwhelmingly lucky to be a collaborator, roommate, and friend with The Nice Reid, who has taught me so much about life, research, this country, and beyond. Those lunch, dinner, workouts, and movie nights that we have shared together and every smile, laugh, and inspiration that came from them are unforgettable memories in my life. My debt to him is simply non-deterministic polynomial time solvable—hopefully P $\neq$ NP.

I would like to thank my internship mentors—Hao at Microsoft Research, Francesco and Fabio at IBM Research, and Ananthram at Army Research Laboratory—for hosting me for three beautiful and rewarding summers. Working with Hao has significantly influenced the ways in which I looked at research and definitely empowered me to achieve more in this journey. I would also like to thank my WSDM Doctoral Consortium mentors Lada Adamic and Andrew Tomkins for their valuable suggestions and comments! I am also very grateful for the hospitality that Robin Dunbar, Jure Leskovec, Jie, and their SENRG, SNAP, and KEG group members gave me when I visited them.

Thank you to all my wonderful lab mates at iCeNSA, with whom I have spent my everyday office time: Everaldo Aguiar, Daniel Barabási, Dipa Dasgupta, Louis Faust, Keith Feldman, Chao Huang, Reid, Saurabh Nagrecha, Aastha Nigam, Yihui Ren, Pingjie Tang, Melinda Varga, Xian Wu, Jian Xu, and Yang Yang, as well as people from the CSE department. I would also like to extend thanks to Jasmine Botello and Joyce Yeats for their continuous and patient help on many administrative processes and paperwork over years. Thank you to everyone that helped and encouraged me during my job search process: Everaldo, Bin Bi, Nitesh, Ying Ding, Xiaowen Dong, Tian Jiang, Reid, Omar, Jure, Hao, Xiaolin Shi, Chenhao Tan, Jie, Zoltán, Hao Wang, Sen, and Xiaolin Zhuo.

For my leisure time at Notre Dame, I would like to thank all my friends with whom I have shared both joy and sadness together: Tian, Reid, Jian, Haoyun, Haipeng, Keith, Han, Yijing, Yihui, Ryoko, and Anne! Go Irish!

I owe the biggest debt of gratitude to my parents for their unconditional love and endless support. Special thanks to brave Yeki, who had me at "hello" eleven years ago and has completed me for the days since then and the days that are yet to come.

# CHAPTER 1

# INTRODUCTION

Network Science is defined as "the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena" by the United States National Research Council [159]. The network science narrative dates back to 1736, when Leonhard Euler introduced the notion of Seven Bridges of Königsberg problem [188], which led to the formation of graph theory [18]. Since then, the study of complex networked systems has largely benefited from the mathematical power of graph theory, which offers a formal way to model networks [13].

By harnessing this mathematical power, network science has emerged as a foundational construct through which to understand complex systems by representing and modeling the different scales and modalities of interactions among components of a system. For example, networks have been leveraged to model the connections between individuals in the physical and virtual worlds—referred to as social networks—as well as the (digital) interactions through which information flows—referred to as information networks. While social and information networks have existed throughout the course of human history, the scientific investigation of these networks is relatively recent. Arguably, the foundation of social networks was laid in the 1930s, when Moreno presented the concept of the 'sociogram', a graphical representation of the social structure among elementary school students [155, 215]. An overview of the history of social and information network science—rooted in graph theory and statistical physics, cultivated by social science, and now flourishing in the era of big network data and computational advances—is summarized in Figure 1.1.

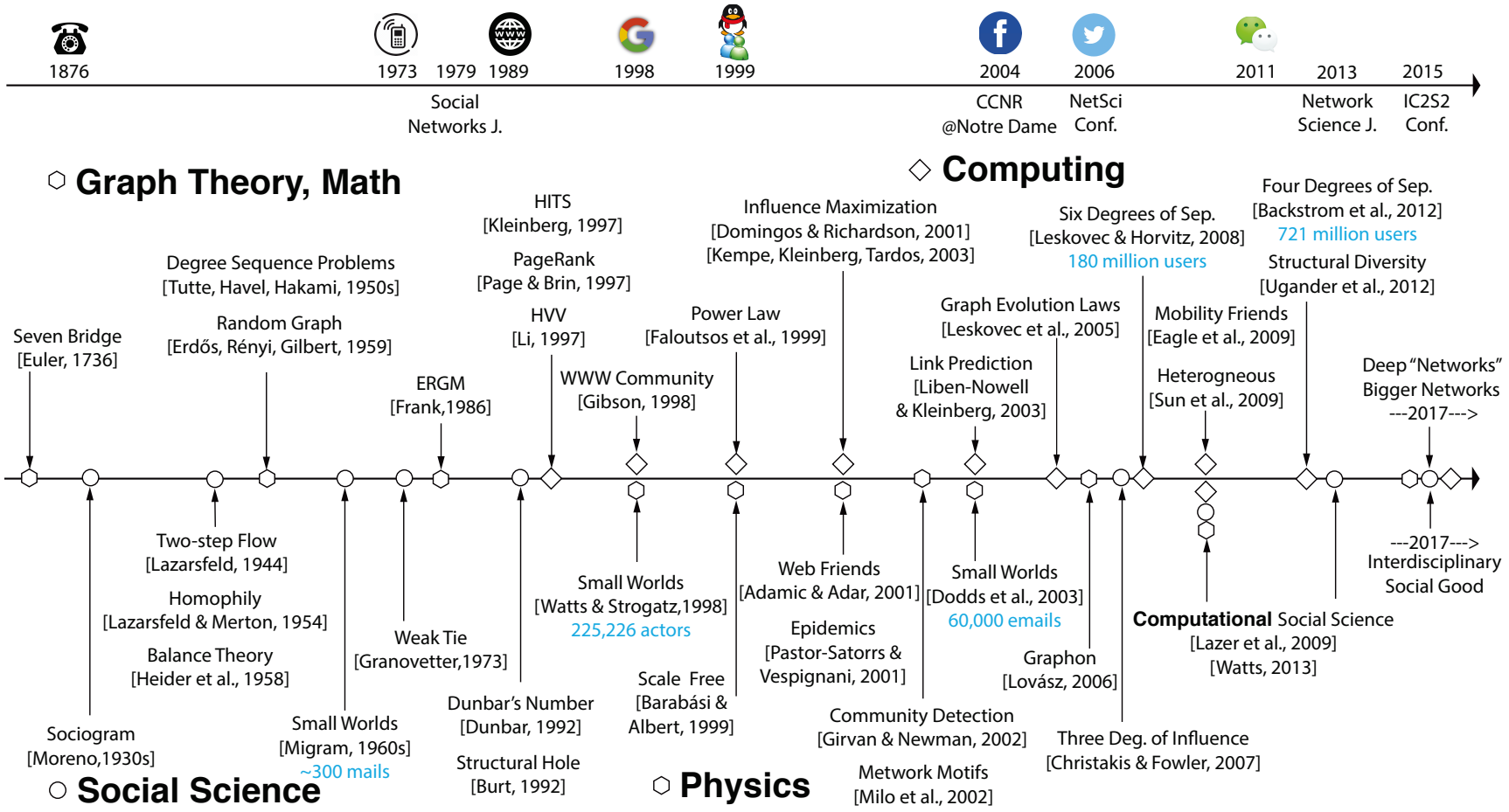



Figure 1.1. A timeline of social and information network science from the perspectives of theory, social science, and computation. The blue text shows the increasing size of networks used in small-world research. The top timeline shows the invention year of several milestone networking services, including the (mobile) phone and World Wide Web, and the inaugural year of several well-known scientific venues on (social) networks.

In the 1950s, abstracted from the $f$-factor problem [213, 214], the degree sequence problem—the problem of whether there exists a graph realization given a degree sequence—started to attract significant interest and effort from mathematicians such as Havel [85] and Hakimi [83]. At the same time, Erdös, Rényi, and Gilbert developed the theory of random graphs [60, 73]—the probability distribution over graphs—and presented the very first random graph model, known as the Erdös-Rényi (ER) model, in which the existence of edges is independent and identical when generating a graph. After their inception, the degree sequence and random graph problems saw extensive exploration by researchers throughout the second half of the $20^{th}$ century, laying the theoretical foundations for (social) network analysis, as well as the grounding for different graph or network science models going beyond the random distribution assumption. Frank and Strauss presented the exponential random graph models (ERGM) [65] in 1986. The Barabási-Albert model, introduced in 1999, proposed the application of the preferential attachment process to generate random networks with the scale-free property [14], a property commonly observed in many networked systems, including social and information networks [14, 62]. By formulating the concept of gradient networks, Toroczkai and Bassler later showed that scale-free networks are emergent from the efficient flow processing (e.g., information) across structures with power-law degree distributions [211]. The Watts-Strogatz model [228], introduced in 1998, enables the reproduction of random networks with two small-world properties—short average path lengths and high clustering coefficients—which are observed in real-world social networks. The remarkable contribution of this model lies in its pioneering capacity to capture both of these small-world properties, neither of which was recoverable by earlier models.

Parallel to the combinatorial and mathematical approaches to network science, there has been a strong impetus from sociological theories. For example, the "small world" phenomenon in social networks was first actualized and documented in a se-

ries of classic mail-tracing social experiments (*296* and *160* letters) conducted by Travers and Milgram in the 1960s [147, 212]. They found that any one individual may be capable of reaching any other via a relatively short chain of intermediaries. Lazarsfeld and Katz proposed the theory of a "two-step flow of communication" to characterize the diffusion process by which information or influence flows from opinion leaders [100, 116]. Lazarsfeld and Merton also established the principle of homophily in 1954, which suggests that individuals generally gravitate to associate with similar others [115]. Furthermore, Heider et al. presented the balance theory to abstract the "friends' friends are friends" and "enemies' enemies are also friends" phenomena, whereby balance is achieved in a closed triangle when all three links or only one link are positive [86]. Both the theories of homophily and social balance imply that close friends have the tendency to have overlapping social circles. However, observing in a 1960s' interview population that relatively weak acquaintances are actually more likely to help job seekers find jobs, Granovetter discovered "the strength of weak ties", the theory that weak ties serve as the bridge between two close people [80]. Granovetter also proposed the concept of embeddedness—"the extent that a dyad's mutual contacts are connected to one another"—to measure the closeness of two people [79, 81]. More recently, in 1992, Ronald Burt formalized the theory of structural holes, according to which those who serve as bridges are able to facilitate the diffusion of information and innovations by bridging the gap between individuals or parties who have complementary sources to knowledge [24].

Combining the foundational aspects from mathematical and social sciences with the surge of digital data and computing advances has led to the development and rapid growth of social and information network research from a computational perspective since the end of the $20^{th}$ century. In particular, the underlying computing challenges resulting from the unprecedented scale of network data started to draw substantial attempts and effort from computing scientists. For example, instead of man-

ually tracing hundreds of letters like the small-world experiments of the 1960s, Watts and Strogatz re-examined the small number of degrees of separation in a *225,226*-actor collaboration network (IMDb) in 1998 [228]. In 2003, Dodds, Muhamad, and Watts used a *60,000* email-tracing replication on a cross-nationally diverse population, producing results encouragingly close to those of the original study [41]. During the same period, Kleinberg also contributed to this line of research from an algorithmic and computational perspective [103–105]. More recently, the existence of the small-world phenomenon has been successfully established using observational data obtained digitally from societal-scale systems featuring millions of individuals and billions of connections, such as *180 million* users in Microsoft MSN Messenger network [120] and *721 million* users in Facebook [12], all enabled by the sheer processing power of modern computers. Furthermore, the computational analysis in large-scale social systems also offers the opportunity to experimentally verify or reject classical conjectures and theories. For example, an examination of the social recruitment process in Facebook by Ugander et al. found that the contagion rate is tightly influenced by the diversity of an individual's neighborhood, upending the conventional wisdom that such rates are controlled by the size of the neighborhood [216].

In addition to the validation of social theories previously established by small-scale field studies, computational lens on big networks also enable the identification and rectification of problems at the societal scale—problems that could not possibly be examined otherwise. In the late 1990s, the Web page ranking problem in information networks was identified and several extraordinary link analysis algorithms were proposed to solve it, including Kleinberg's Hyperlink-Induced Topic Search (HITS) [106], Page et al.' PageRank [23], and Li's Hyperlink Vector Voting (HVV) [124]. By building upon the concepts of authorities and hubs in HITS, Gibson et al. defined and inferred the community structures on the WWW, and derived possible explanations for the grouped and hierarchical organization of the WWW network [72]. The early

part of the $21^{st}$ century has seen explosive growth in the exploration and study of social and information networks using computational tools and perspectives. Between 2001 and 2003, the influence maximization problem in social networks was proposed and formalized as the NP-hard optimization problem of selecting at most $k$ users who are able to maximize the 'word of mouth' effect in a social network [42, 101]. Meanwhile, Liben-Nowell and Kleinberg formalized the question of whether new links between individuals are predictable from a previous snapshot of a social network as the link prediction problem [125]. Subsequently, Leskovec et al. studied the problem of network evolution and found many social and information networks follow the densification laws and exhibit shrinking diameters over time [121]. Undoubtedly, over the last two decades, a myriad of remarkable problems, discoveries, and applications have been demonstrated by applying a computational lens to large-scale social and information networks [57], leading to the emergence of an exciting field: computational social science [117, 226].

Notwithstanding the cornucopia of substantial and influential work already achieved in network science and computational social science, there remain various unanswered questions, some of which are the focus of this dissertation. First, the interplay in human social networks between the rich set of demographic traits associated with individuals and the underlying network structures is still poorly understood. There is evidence concerning the role of demographics in social activities, such as the principle of (demographic) homophily that facilitates the connecting and maintaining of relationships [140] and the effects of cultural tastes on the density of social contacts across social ties with different tie strength [127]. However, if we consider the influential scientific milestones discussed above, many questions remain unexplored, especially as to how demographic properties inform network structuration. Specifically:

- Regarding the theory of weak/strong ties, what influence do the gender and age of two people connecting with a social tie have on its strength?

- With respect to the social balance theory, how is balance achieved among three people of different gender, such as three females or three males?
- Concerning the small world phenomenon, do individuals in specific age groups live in a small world in relation to individuals in the same or other generational clusters?
- With these social theories developed, are user demographic profiles predictable from network structures?
- To untangle the interrelations between individuals' demographics and connections, how do we address the computational challenges that arise from the unprecedented scale of big network data?

In addition, there is also a lack of understanding of networks when coupled with the notion of diversity. Research has shown that diversity matters in a wide range of systems and disciplines. In biology and genetics, the stability of an ecosystem benefits from the variability of species or genetics within it [66]. In organizational and economic science, the diversity level of a group is more likely to positively correlate with its performance [169]. More recently, network and social scientists have studied the effects of structural diversity on economic development [56] and information contagion [216]. Herein, we are interested in the following unaddressed questions relating to when diversity meets networks:

- In view of the principle of (link) homophily, where individuals with more friends in common are more likely to associate with each other, how does the structural diversity of common neighborhoods influence the likelihood that an association exists or will form between a pair of individuals?
- As we are embedded in numerous networks, does the influence of common neighborhood diversity vary across different social and information networks?
- At the individual level, how does the topical diversity and authority of a scientist's research affect the growth of his or her scientific impact?
- Networks composed of diverse types of objects and connections present unique challenges that cannot be handled by conventional models for homogeneous networks. How do we design computational models to incorporate the aspects of network diversity?

Answers to these demographics- and diversity-based questions are not only critical to furthering the development of the network and social sciences, but also to better understanding the dynamics of human behavior, and the formation and sustenance of societies. In light of this, we develop computational models to answer these questions in large-scale networked social and information systems, and also demonstrate various applications of the same. This research follows the process from discovering social science inspired mechanisms to descriptive analytics to predictive science.

## 1.1 Contributions and Organization

In this thesis, we harness the power of social, data, and network science to unveil the social phenomena that emerge from individuals' interactions, and ultimately model and predict user behavior in large-scale social and information networks. Our research is based on a rich collection of big network data that are digitally recorded, from mobile phone communications to online social media to digital libraries. Using this large collection of big network data, we formalize a wide range of data and network science problems and identify their unique challenges in the context of networks. We present computational models to address these problems and to efficiently and effectively extract knowledge from these massive network datasets. The final thrust of our research is then devoted to translating these discoveries and computational models into real-world applications and insights.

As mentioned above, this thesis focuses on two broad and important directions—demographics and diversity—and the way in which they manifest in big networks. This enables a natural separation into two parts, one dedicated to each direction. The overview of this thesis is shown in Figure 1.2. In Part I, we study the significant social strategies that are used by females and males to fulfill their social needs—one of the basic human needs—across their lifespan. Second, we investigate the phenomenon of demographic-specific small worlds, and quantify the differences in the "smallness" of

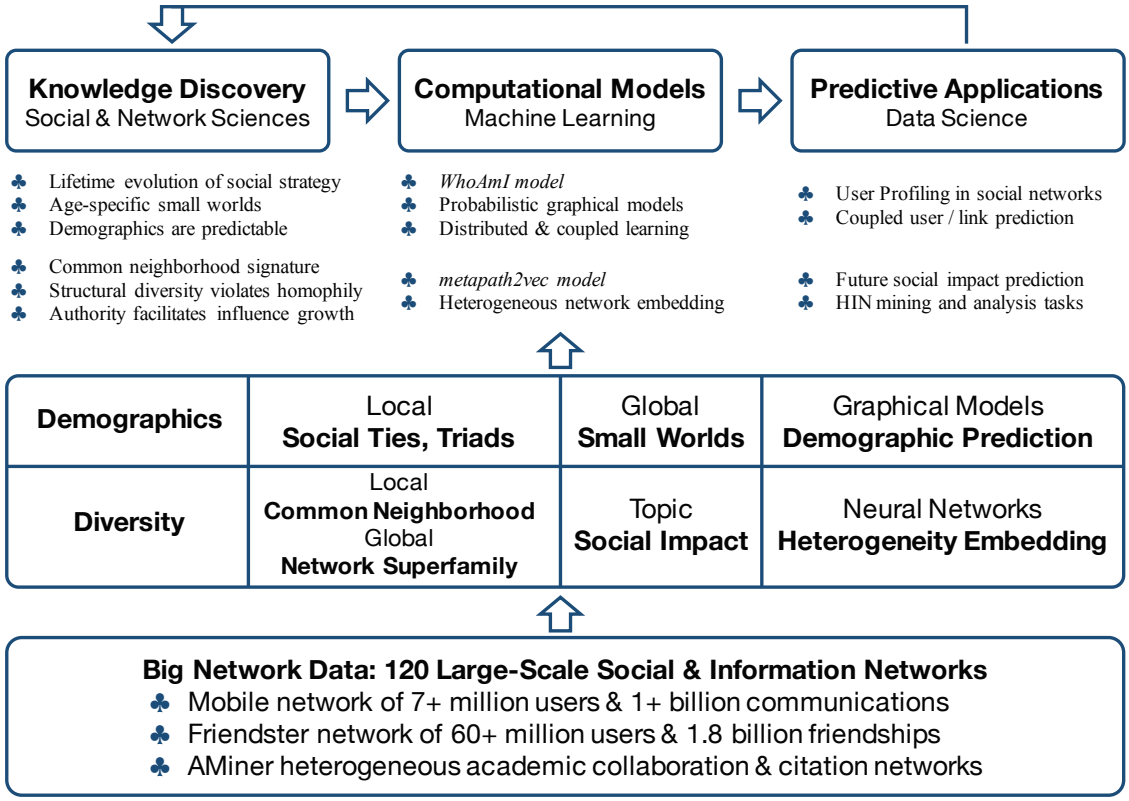



Figure 1.2. The structure of the thesis. Given the large-scale social and information network data at the bottom, the topics that this thesis investigates are presented in the middle and the flow of the scientific process this thesis follows is shown at the top, wherein the major discoveries and contributions are summarized at each step.

the world in which individuals of different gender and age live. From a computational perspective, we formalize the problem of the joint inference of multiple demographic attributes across coupled networks and present an effective and efficient learning model to tackle its underlying challenges. Correspondingly, the contributions of Part I are organized into three chapters.

Chapter 2 presents the study of how individuals with different demographic profiles connect and interact with each other in a big mobile communication network of 7 million users and 1 billion phone call and text message records approximating interaction patterns at a societal scale. We investigate the interplay of communication

interactions and demographic characteristics in the perspective of three micro-level network structures—ego networks, social ties, and social triads. We draw a comparison between the usage of phone calls and text messages to network with others. We also demonstrate how the evolution of social strategies used by females and males are dissimilar. We find, first, that young people put greater focus on enlarging social circles; as they age, they have the tendency to maintain small but closed social relationships. Second, we also observe that same-gender triadic relationships are persistently maintained over a lifetime, while the triangles among three opposite-gender individuals disappear as one enters middle-age. Finally, the presented null model demonstrates the statistical significance of the evolving social strategies in human communication.

Chapter 3 presents the investigation of the age-specific small worlds using data from the aforementioned large-scale mobile communication network. Rather than asking whether two random individuals are separated by a small number of links, we ask whether individuals in specific age groups live in a small world in relation to individuals from other age groups. Our analysis shows that there is systematic variation in this age-relative small-world effect. We find young people live in the "smallest world," being separated from other young people and their parents' generation via a smaller number of intermediaries than older individuals. We also discover that the most elderly live in the "least small world," being separated from their same-age peers and their younger counterparts by a larger number of intermediaries. Finally, we demonstrate that variations in the small-world effect are specific to age as a node attribute (being absent in the case of gender) and are consistently observed under several data robustness checks.

Chapter 4 studies to what extent users' demographic profiles can be inferred from their mobile communication patterns, observed from previous chapters. We formalize the demographic prediction problem of inferring users' gender and age si-

multaneously. We propose a factor graph-based *WhoAmI* method to address the problem by leveraging not only the correlations between network features and users' gender/age, but also the interrelations between gender and age. Additionally, we identify a new problem—coupled network demographic prediction across multiple mobile operators—and present a coupled variant of the *WhoAmI* method to address its unique challenges. Our extensive experiments demonstrate the effectiveness, scalability, and applicability of the *WhoAmI* methods. Finally, our study finds a greater than 80% potential predictability for inferring users' gender from phone call behavior and 73% for users' age from text messaging interactions.

In Part II, we investigate the effect of diversity on big networks. First, we study the influence of the diverse structures embedded in the common neighborhood on link formation across more than one hundred large-scale social and information networks. Second, we formalize a novel scientific impact prediction problem to examine factors—topic diversity, authority, popularity, and so on—that can drive a paper to increase its authors' $h$-indices. Finally, we define the problem of heterogeneous network representation learning and present neural network-based models to embed networks of diverse types of nodes and links. This part is naturally structured into the following three chapters.

Chapter 5 examines the principle of structural homophily—people with more common neighbors are more likely to connect with each other—and characterizes the structure of common neighborhoods as a function of their diversity and embeddedness. Using a collection of 120 large-scale networks—the biggest with over 60 million nodes and 1.8 billion edges—we then leverage these structural characteristics to develop a unique network signature, which we use to uncover several distinct network superfamilies not discoverable by conventional methods. We demonstrate that the impact of the common neighbor subgraph on link existence is significantly different across various networks, with its diversity demonstrating a positive effect in BlogCat-

alog and LinkedIn and a negative effect in Facebook and Friendster. We also discover striking cases where it violates the principle of homophily. Our findings suggest that the common neighborhood signature (CNS) is an intrinsic network property.

Chapter 6 presents work that aims to answer the question of whether a scientific publication will contribute to its authors' future $h$-indices, in contrast to the traditional focus on predicting the exact citation count of this publication in a regression fashion. Using the AMiner dataset with millions of authors and papers, we find that the researcher's authority on the publication topic and the venue in which the paper is published are crucial factors to the increase of the primary author's $h$-index, while the topic diversity and popularity are of surprisingly little relevance. By leveraging relevant factors, we can predict an author's $h$-index in five years with an $R^2$ value of 0.92 and whether a previously (newly) published paper will contribute to this future $h$-index with an $F_1$ score of 0.99 (0.77). Finally, we develop an online tool that allows users to generate informed $h$-index predictions.

Chapter 7 investigates how neural network-based embedding models can advance heterogeneous network mining and analysis. We begin by formally defining the heterogeneous network representation learning problem. To address the unique challenges that result from network heterogeneity, we propose the *metapath2vec* and *metapath2vec++* frameworks that are capable of capturing both the structural and semantic correlations of nodes and relations with different types. Extensive experiments demonstrate that the learned latent representations by *metapath2vec* can be applied to various mining tasks in heterogeneous information networks, including similarity search, classification, and clustering. We conclude that properly accommodating these advantages may further improve the mining and learning tasks in heterogeneous information networks.

In Chapter 8, we conclude this thesis and look into the future of big network analytics from a computational perspective.

# PART I

# DEMOGRAPHICS IN BIG NETWORKS

# CHAPTER 2

# GENDER & AGE IN NETWORKS

## 2.1 Overview

In this chapter, we harness the power of network and data sciences to model the interplay between user demographics and social behavior. By studying millions of users and more than one billion mobile communication records, we unveil the significant social strategies that are used by people to satisfy their social needs across the lifespan. Specifically, we investigate the correlations between demographic characteristics and micro-level social structures—ego networks, social ties, and social triads. First, we find young people put more focus on enlarging social circles; as they age, they have the tendency to maintain small but closed social relationships. Second, we also observe striking gender differences in triadic relationships across individuals' lifespan, that is, the connections among three same-gender users are persistently maintained over a lifetime, while the opposite-gender triadic relationships disappear when users enter into their middle-age. Third, we observe frequent cross-generation interactions that are essential for bridging age gaps as well as cross-gender communications, in particular during dating active age, that are important for maintaining romantic relationships. Finally, the present null model demonstrates the statistical significance of the evolution of social strategies in human communication.

This chapter is largely extracted from previous publications [44, 53]. It is a joint work with Jie Tang, Yang Yang (THU), Yang Yang (ND), and Nitesh V. Chawla.

## 2.2 Introduction

As of 2016, the number of mobile users is 4.611 billion, corresponding to a global penetration of 62%; The number of mobile subscriptions across the globe reaches 7.377 billion in 2016, which is approximately the same with the world population, from a recent report by the International Telecommunications Union (ITU). On average, each mobile user makes, receives or avoids 22 phone calls and sends or receives text messages 23 times, and checks their phones up to 150 times a day [192]. These mobile devices record huge amounts of user behavioral data, in particular users' daily communications with others. This provides us with an unprecedented opportunity to study how people build and maintain connections in mobile communication networks.

Previous work on mobile communication networks mainly focused on macro-level models, like network distributions [164], scale free [54], duration distributions [43, 180], and mobility modeling [47, 77, 223]. Recently, researchers have also started to pay more attention to the micro-level analysis of the mobile networks. For example, Eagle et al. [55] studied the friendship network of 100 specific mobile users (students or faculties at MIT). They investigated human interactions (what people do, where they go, and with whom they communicate) based on the machine-sensed environmental data collected by mobile devices. Meng et al. [143] studied the mobile communication networks of 200 students at the University of Notre Dame. They explored the interplay between individuals' evolving interaction patterns and traits. However, these work did not consider the interplay between user demographics and communication behavior. More recently, Nokia Research organized the 2012 Mobile Data Challenge to infer mobile user demographics by using communication records of 200 users [152, 236]. However, the scale of the network is very limited. In this chapter, we leverage a large-scale mobile network to study how users' communication behaviors correlate with their demographic attributes.

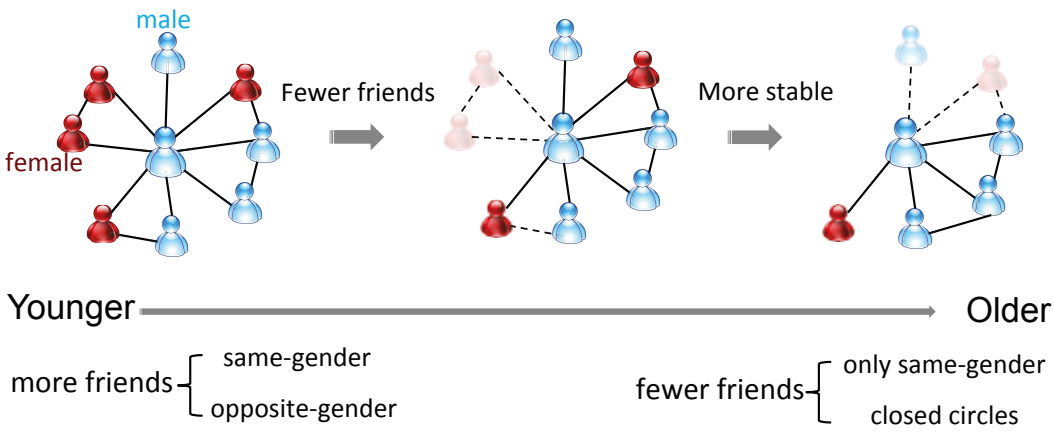



Figure 2.1. Evolution of demographic-based social strategies in human communication.

**Contributions.** In this chapter, we employ a real-world large mobile network comprised of more than 7,000,000 users and over 1,000,000,000 communication records (voice phone call and short text messaging) as the basis of our study, which we use to systematically investigate the interplay of user communication behavior and demographic information. Through the study, we first unveil several intriguing *social strategies* that users of different age and gender use to meet their social needs, i.e., building new connections and maintaining existing relationships. Simultaneously, we examine the differences between people's phone call and text messaging behavior. To the best of our knowledge, we are the first to study the problem of inferring demographic-based social strategies in such a real-world large mobile network.

This chapter investigates social strategies from both the voice phone call network and the short text messaging network, and further conclude the networking differences and similarities between human phone call and text messaging behaviors. In specific, we examine the interplay between user demographics and three different types of micro-network structures, including ego networks, interpersonal ties, and social triads. In addition, we propose to use a null model to validate the statistical significance of social strategies observed from network structures.

**Key Findings.** Our study unveils the significant social strategies and their evolution across the lifespan in human communication, which are highlighted in Figure 2.1. Specifically, we discover that younger people are very active in broadening their social circles, while older people tend to maintain smaller but more closed connections. We find that the communications between two younger opposite-gender users are more frequent than those between same-gender users. We also observe frequent cross-generation interactions that are essential for bridging age gaps in family, workplace, education, and human society as a whole [141]. We unveil that people expand both same-gender and opposite-gender connections during their active dating period (18 – 34 years old), while they maintain only same-gender social groups in mobile communication after 35 years of age. Finally, our analysis shows strong interrelations between users' age and gender. For example, a 20-year-old female's social networking behavior is distinct from not only a 20-year-old male's, but also from a 50-year-old female's.

## 2.3 Mobile Network Data with Demographics

The dataset used in this chapter is extracted from a collection of more than 1 billion (1,000,229,603) phone call and text messaging events from an anonymous country [44, 49, 58, 77], which spans from Aug. 2008 to Sep. 2008. Notice that we only consider the communications that were made between users within this country. We construct two undirected and weighted mobile communication networks from the de-identified and anonymous data: a phone call network (referred to as CALL) and a text messaging network (referred to as SMS). To represent the human communication behavior in networks, we place an edge between two users if and only if they have reciprocal communications (voice calls or text messages) within the observation timeframe [164]. Specifically, we view each user as a node $v_i$ and create an edge $e_{ij}$ between two users $v_i$ and $v_j$ if and only if they made reciprocal calls or text messages

TABLE 2.1

THE STATISTICS OF MOBILE NETWORKS

| networks | #nodes | #edges |
|---|---|---|
| CALL network with user demographics ($\text{CALL}_d$) | 7,440,123 | 32,445,941 |
| SMS network with user demographics ($\text{SMS}_d$) | 4,505,958 | 10,913,601 |
| Reciprocal CALL network ($\text{CALL}_r$) | 4,927,095 | 16,674,164 |
| Reciprocal SMS network ($\text{SMS}_r$) | 3,104,853 | 7,602,830 |
| Largest Connected Component of $\text{CALL}_r$ ($\text{CALL}_{rl}$) | 4,295,638 | 15,787,538 |
| Largest Connected Component of $\text{SMS}_r$ ($\text{SMS}_{rl}$) | 2,369,078 | 6,660,172 |
| $\text{CALL}_{rl}$ with user demographics ($\text{CALL}_{rld}$ / CALL) | 4,292,227 | 15,765,196 |
| $\text{SMS}_{rl}$ with user demographics ($\text{SMS}_{rld}$ / SMS) | 2,064,898 | 5,689,696 |

($v_i$ called $v_j$ and also $v_j$ called $v_i$ for at least one time during the observation period). The strength $w_{ij}$ of the edge is defined as the number of communications between $v_i$ and $v_j$ per month. Then we extract the largest connected component from each network as our experimental networks. We also generate the networks by filtering out the nodes that don't have demographic information. Table 2.1 lists the order and size of the resultant CALL and SMS networks. The data does not contain any communication content.

In this dataset, around 45% of the users are female and 55% are male. We compare the demographic population distribution of mobile users with the 2008 world population distribution, which was released by the U.S. Census Bureau international database[1]. We find that both female and male users between the ages of 20 and 55 are strongly overrepresented in the mobile population compared to the global

[1]U.S. Census Bureau. `http://www.census.gov/idb/worldpopinfo.html`. Jan. 1st, 2014

TABLE 2.2

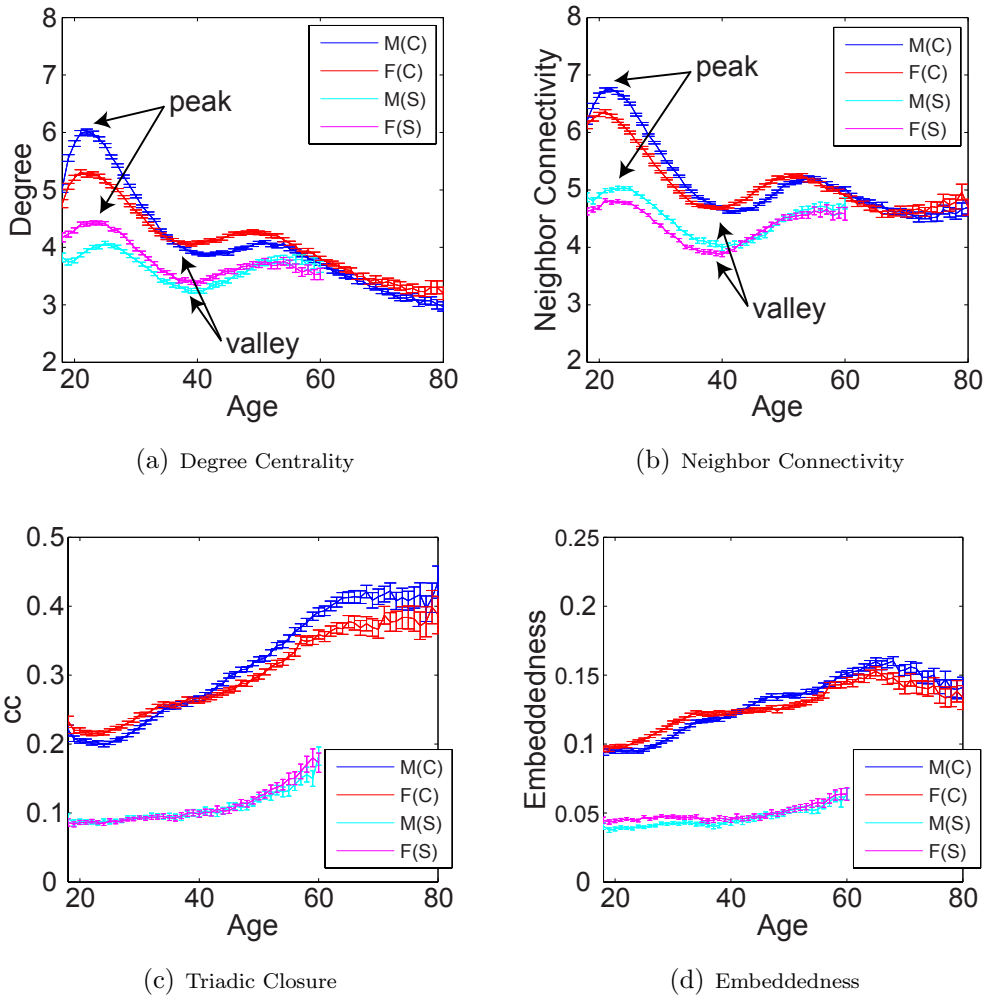
THE DISTRIBUTION OF MOBILE USERS' GENDER AND AGE

| | Young | Young-Adult | Middle-Age | Senior |
|---|---|---|---|---|
| Female | 4.77% | 13.52% | 16.16% | 10.84% |
| Male | 5.23% | 15.96% | 19.73% | 13.66% |

population, while teenagers (under 18 years old) and the elderly (aged 80 or over) are underrepresented. Thus in our study, we focus on users aged between 18 and 80 years old. To simplify the notations, we use $F$ and $M$ to denote the female and male users, respectively. Following [17, 95], we also split users into four groups according to their ages: Young (18 – 24), Young-Adult (25 – 34), Middle-Age (35 – 49), and Senior ($> 49$). The distribution of users' gender and age is listed in Table 2.2.

## 2.4 Social Strategies in Mobile Communication

Social strategies are used by people to meet their social needs that is, together with being, having, and doing, considered among the basic human needs [138]. Meeting with new people and strengthening existing relationships belong to the category of social needs. The mobile communication data provides rich information for discovering and characterizing human social strategies by which people build and maintain social connections. Previous studies [166] show that the strategies by which social needs are satisfied change over time, although the needs are constant across one's lifetime. In this section, we show how people communicate with each other across their respective lifetime. Specifically, we investigate the interplay of human communication interactions and demographic characteristics in the perspective of micro-level network structures, including ego networks, social ties, and social triads. We also use




(a) Degree Centrality

(b) Neighbor Connectivity

(c) Triadic Closure

(d) Embeddedness

Figure 2.2. Correlations between demographics and network characteristics. $C$ means attributes observed from the CALL network and $S$ means the SMS network. $F$ denotes female and $M$ denotes male.

a null model to simulate the observations by randomly shuffling users' demographic profiles and report the statistical significance of the results in Section 2.5.

### 2.4.1 Social Strategies on Ego Networks

An ego network of one person is defined by viewing himself or herself as the central node and his or her one-degree friends as surrounding nodes [67]. Clearly, one's ego network is a sub-network of the original network. Figure 2.1 presents an illustrative

example of the evolution of one's ego network. We first examine the characteristics of the central node (ego) and then the distributions of this ego's friends (ego network) with respect to their demographic profiles.

**Ego.** We present a basic correlation analysis between network characteristics and user demographics to examine how an individual's gender and age influence her or his ego social networks. In particular, we consider the following network metrics:

- *Degree Centrality*: the number of edges incident upon a node in the network;
- *Neighbor Connectivity*: the average degree of neighbors of a specific user.
- *Triadic Closure*: the local clustering coefficient (*cc*) of each user;
- *Embeddedness*: the degree that people are enmeshed in networks [81]. More accurately, a user $u$'s embeddedness is defined as $\frac{1}{|N_u|}\sum_{v\in N_u}\frac{|N_u\cap N_v|}{|N_u\cup N_v|}$ , where $N_u$ is the neighbors of $u$.

Figure 2.2 plots the correlations between the four network metrics and the users' age. From sub-figures 2.2(a) – 2.2(b), we observe that the degree and neighbor connectivity of both female and male users achieve peak values around 22 years old, then decrease with valleys around 38 – 40 years old. An interesting phenomenon is that before this valley, the males have clearly higher scores on both metrics (degree and neighbor connectivity), while the situation is reversed after this point.

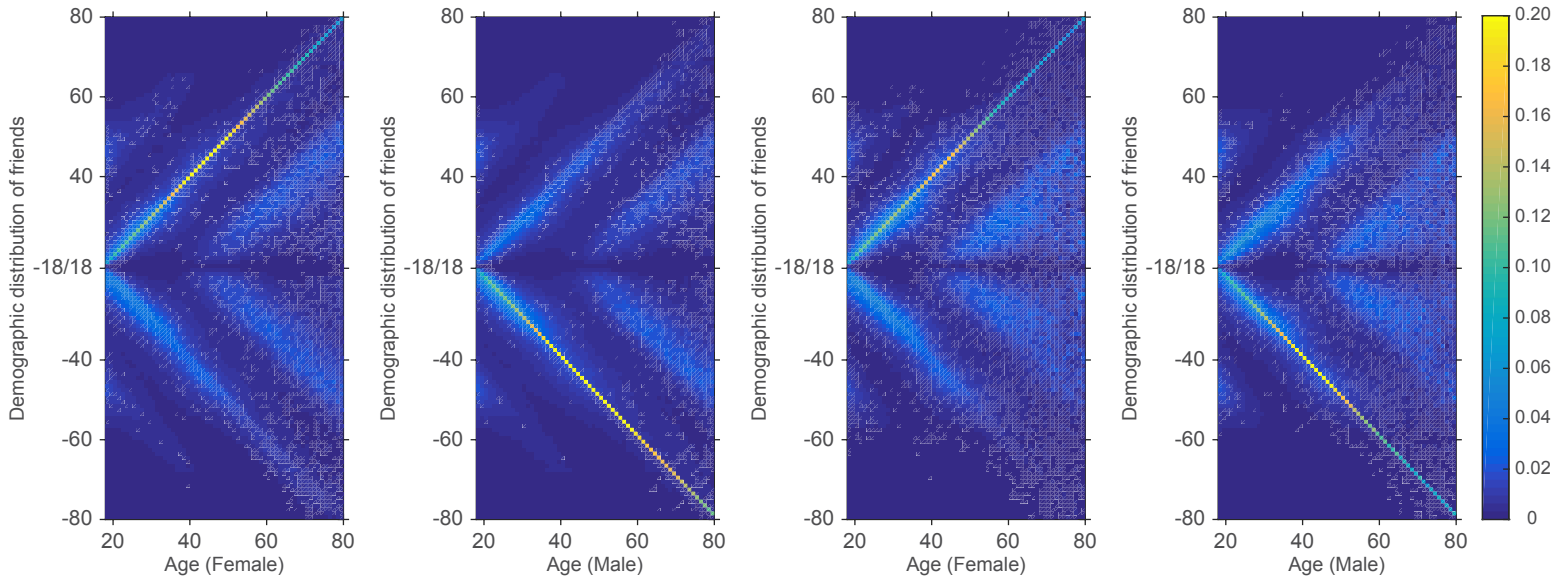
From sub-figures 2.2(c) – 2.2(d), we see that both triadic closure and embeddedness increase when users become older. Similar to the first two metrics, there is also a reverse phenomenon at age 38 – 40. The difference lies in that the male's triadic closure and embeddedness are at first smaller than the female's, and then become larger after the reversion point. All four network metrics are observed at a 95% confidence interval.

**Ego Networks.** With one's ego network, we study the demographic homophily on both gender and age. The principle of homophily suggests that people tend to be connected with those who are similar to them [115]. It has been extensively studied

and verified in both online social networks [120, 130] and mobile networks [43, 110].

Figure 2.3 shows friends' demographic distribution for female and male users of different age in the CALL and SMS networks. The $x$-axis represents a central user' age from 18 to 80 years old and the $y$-axis represents the demographic distribution of that central user' friends, in which positive numbers denote female friends' age and negative numbers denote male friends'. The spectrum color, which extends from dark blue (low) to yellow (high), represents the probability of one's friends belonging to the corresponding age ($y$-axis) and gender (positive or negative). Interestingly, there exist highlighted diagonal lines in each sub-figure, which suggests that people tend to communicate with others of similar age. In particular, the age homophily is much stronger for people aged between 35 to 55 years old in the CALL network, and 40 to 50 years old in the SMS network. Simultaneously, the highlighted diagonals appear in the same gender range in both networks, i.e. females appear in the positive Y range (F) in Figures 2.3(a), 2.3(c) and males in the negative Y range (M) in Figures 2.3(b), 2.3(d), which shows the existence of a high degree of gender homophily in mobile phone behavior.

**Social Strategies.** From a sociological perspective, the results in Figures 2.2 and 2.3 can be also explained by different social strategies that people use to maintain their social connections. First, younger people (who have higher degree centrality) are very active in broadening their social circles, while older people (who have higher triadic closure centrality $cc$) tend to keep smaller but more stable connections. This finding from large-scale networks coincides with previous survey studies that older people have lower rates of contact than young people [35, 136]. Second, people tend to communicate with others of similar gender and age, i.e., gender and age homophily in mobile communications. Third, young people put increasing focus on the same generation and decreasing focus on the older generation, and the middle-age people devote more attention on the younger generation even at the cost of age homophily.




(a) F ego net in CALL (b) M ego net in CALL (c) F ego net in SMS (d) M ego net in SMS

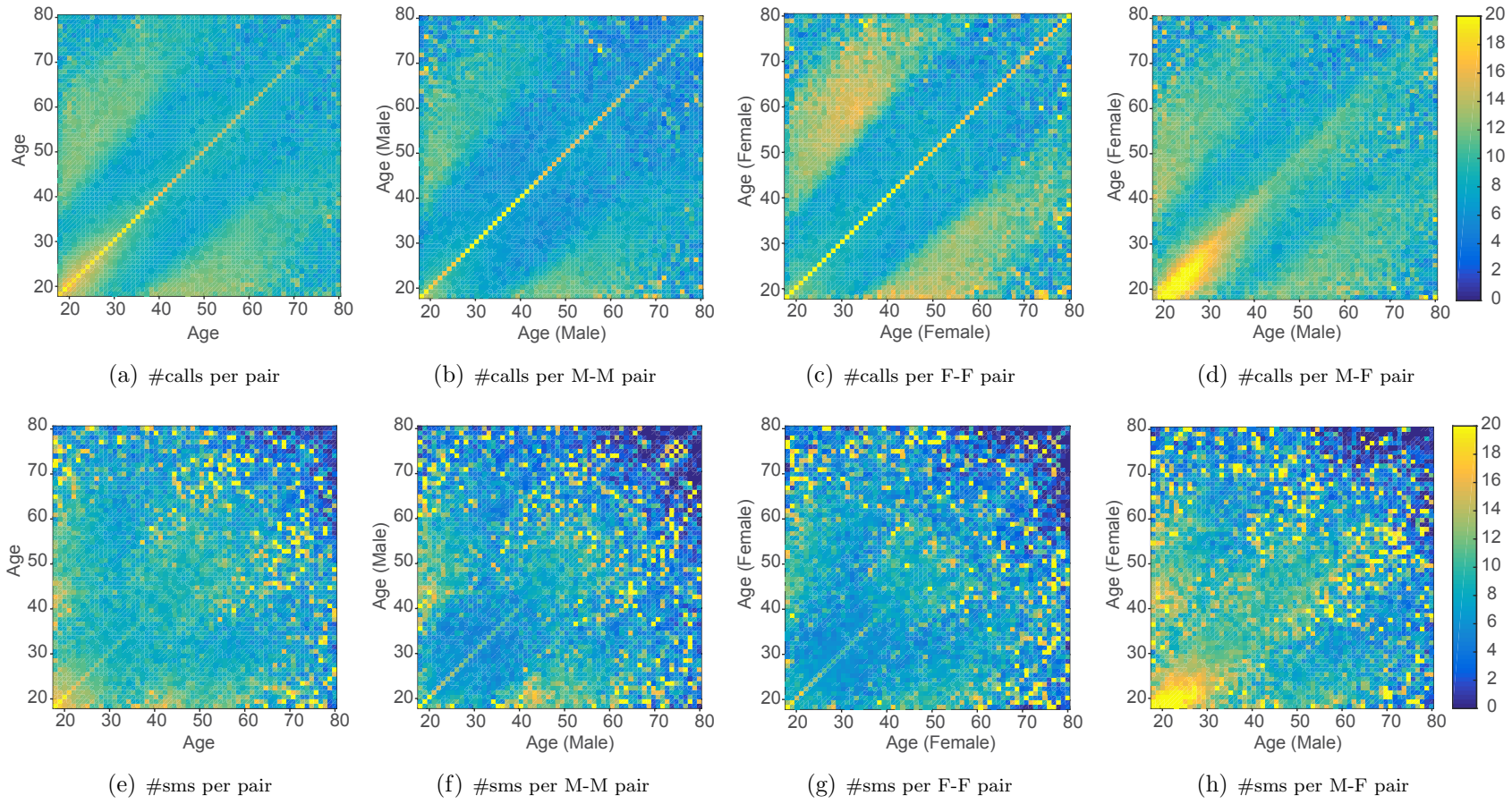
Figure 2.3. Friends' demographic distribution in ego networks. $x$-axis: (a) the age of a female ego in CALL; (b) the age of male ego in CALL; (c) the age of a female ego in SMS; (d) the age of a male ego in SMS. $y$-axis: the age of the ego's friends (positive: female friends, negative: male friends). The spectrum color represents the friends' demographic distribution.

### 2.4.2 Social Strategies on Interpersonal Ties

An interpersonal tie is viewed as the connection between two people, and its strength represents the extent of closeness of social contacts [164], such as strong ties [111, 183] and weak ties [80]. In mobile communication networks, tie strength is defined as the frequency of communications between each pair of users [164, 166].

In Figure 2.4, we use heat maps to visualize the communication frequencies for different demographics. Figures 2.4(a) and 2.4(e) report the average number of calls/messages per month between two users. Figures 2.4(b) – 2.4(d) and 2.4(f) – 2.4(h) detail the analysis by reporting the average numbers of calls/messages between two male users, two female users, and one male and one female, respectively. Again, we discover highlighted diagonal lines in Figures 2.4(a) – 2.4(c), which correspond to the gender and age homophily. We also notice that there are highlighted areas corresponding to cross-generation communications. In Figure 2.4(a), the color of cross-generation areas that extends from green to yellow indicates that on average 13 calls per month have been made between people aged 20 – 30 and those aged 40 – 50 years old. This potentially corresponds to phone calls between parents and children, managers and subordinates, and advisors and advisees, etc. These two discoveries can also be observed in Figures 2.4(e) – 2.4(g) in the SMS network but not as obvious as in the CALL network.

In addition, we observe that the cross-generation phone call communications between female users seem to be much more frequent than those between male users (Cf. Figures 2.4(b) and 2.4(c)). Moreover, from Figures 2.4(d) and 2.4(h), we observe a highlighted yellow area between people aged 18-34 years old, which means that cross-gender communications are more frequent than those between users of the same gender. A similar observation has also been reported in the MSN network [120].




(a) #calls per pair (b) #calls per M-M pair (c) #calls per F-F pair (d) #calls per M-F pair

(e) #sms per pair (f) #sms per M-M pair (g) #sms per F-F pair (h) #sms per M-F pair

Figure 2.4. Strength of social ties in the CALL and SMS networks. $x$- and $y$-axis: age of users with specific gender. The spectrum color represents the number of phone calls (text messages) per month. (a), (b), (c), (e), (f), and (g) are symmetric.

**Social Strategies.** The social strategies unveiled from Figure 2.4 can be summarized as follows. First, frequent cross-generation interactions are maintained to bridge age gaps in both phone call and text messaging channels. Second, opposite-gender communication interactions among younger people are much more frequent than those between same-gender users. However, when people reach the 35 years of age, reversely, same-gender interactions are more frequent than those between opposite-gender users.

### 2.4.3 Social Strategies on Triads

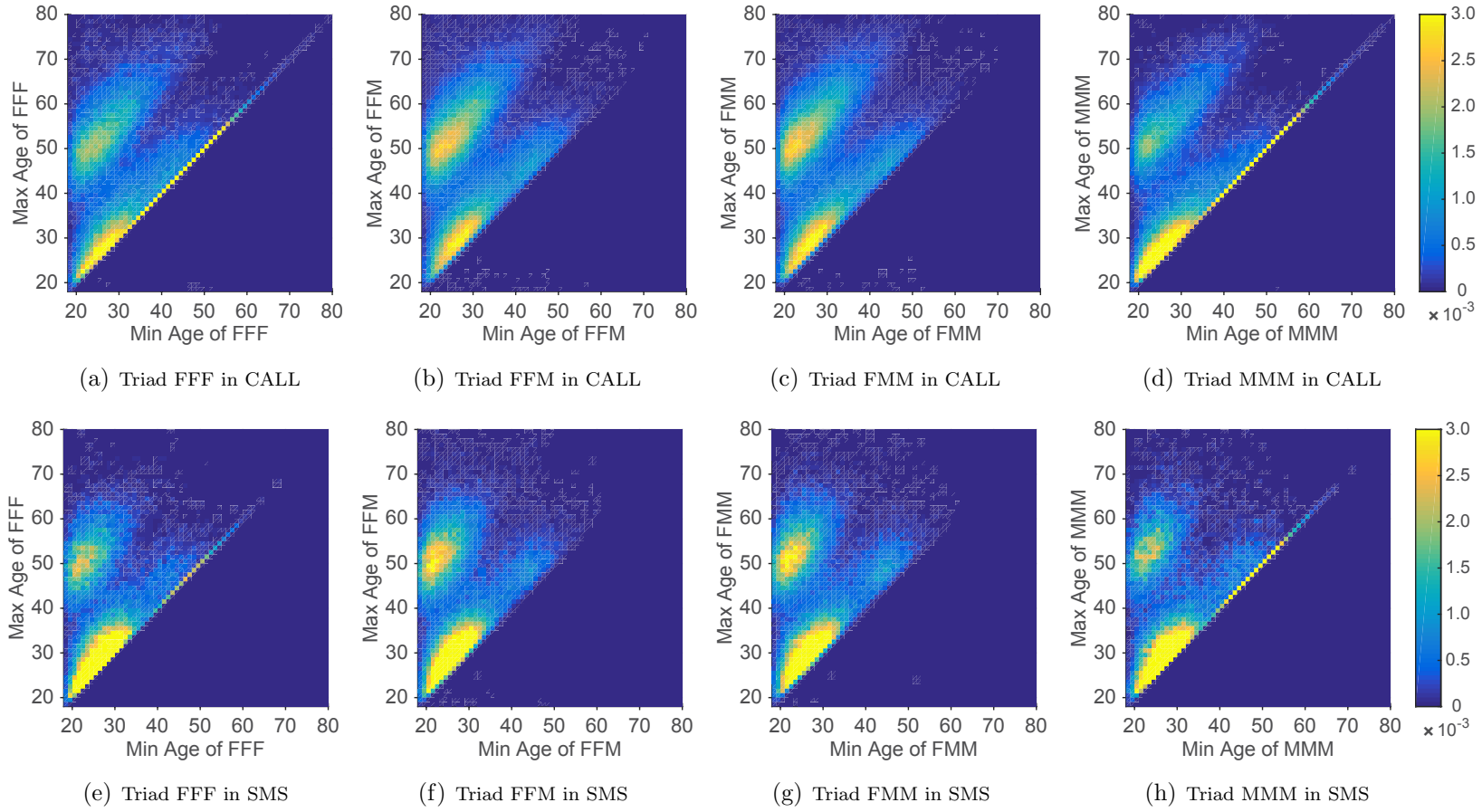
A triad is one of the simplest groupings of individuals in social networks [57]. Three individuals form a triad if and only if each pair of them are friends. Herein, we investigate how male and female users maintain their social triadic relationships across their lifetime.

In Figure 2.5, the heat map visualizes the distribution of the minimum age ($x$-axis) and maximum age ($y$-axis) of three users in a closed social triad structure. Figures 2.5(a)/2.5(e) and 2.5(d)/2.5(h) show the same-gender triads: 'FFF' and 'MMM', and Figures 2.5(b)/2.5(f) and 2.5(c)/2.5(g) present the age distribution for users in opposite-gender triads: 'FFM' and 'FMM'. Clearly, the triadic relationships are observed in all four kinds of gender-triads (i.e., 'FFF', 'MMM', 'FFM' and 'FMM') among young people by highlighted yellow areas at the left-bottom corners of each sub-figure. When entering middle-age ($>$ 35 years old), people only maintain the same-gender triadic relationships in mobile communications, which is revealed by the yellow diagonal lines in Figures 2.5(a)/2.5(e) and 2.5(d)/2.5(h). The opposite-gender triadic relationships vanish when people pass 35 years old observed in Figures 2.5(b)/2.5(f) and 2.5(c)/2.5(g). The instability of opposite-gender triadic relationships and the persistence of same-gender triadic relationships across one's lifetime are novel discoveries and reveal the dynamics of human social strategies across their

lifespan.

Furthermore, the cross-generation triadic relationships are found in the left-middle light areas in each sub-figure. These left-middle light areas are almost isolated with other highlighted areas in each sub-figure, then we are curious about the distribution of the middle age of three users in one social triad. Our further study shows that the middle age in these triads are similar to either the minimum age (60%) or the maximum age (40%) among them, which means there are around 60% cross-generation communication triads are composed of two youths and one middle-age people, for example, 25-25-45 years old respectively in a triad, the remaining 40% are two middle-age and one young people, for example, 20-40-40 years old, and no triads like 20-30-40 years old are observed in this nationwide communication networks.

**Social Strategies.** The dynamics of gender differences on social decisions indicate the evolution of social strategies used by people to meet their social needs. People expand both the same-gender and opposite-gender social circles during the dating active period. However, people's attention to opposite-gender groups quickly disappears after entering into middle-age, and the insistence and social investment on same-gender social groups last for a lifetime.




(a) Triad FFF in CALL (b) Triad FFM in CALL (c) Triad FMM in CALL (d) Triad MMM in CALL

(e) Triad FFF in SMS (f) Triad FFM in SMS (g) Triad FMM in SMS (h) Triad MMM in SMS

Figure 2.5. Social triad distribution in the CALL and SMS networks. $x$-axis: the minimum age of three users in a triad. $y$-axis: the maximum age of three users. The spectrum color represents the distributions.

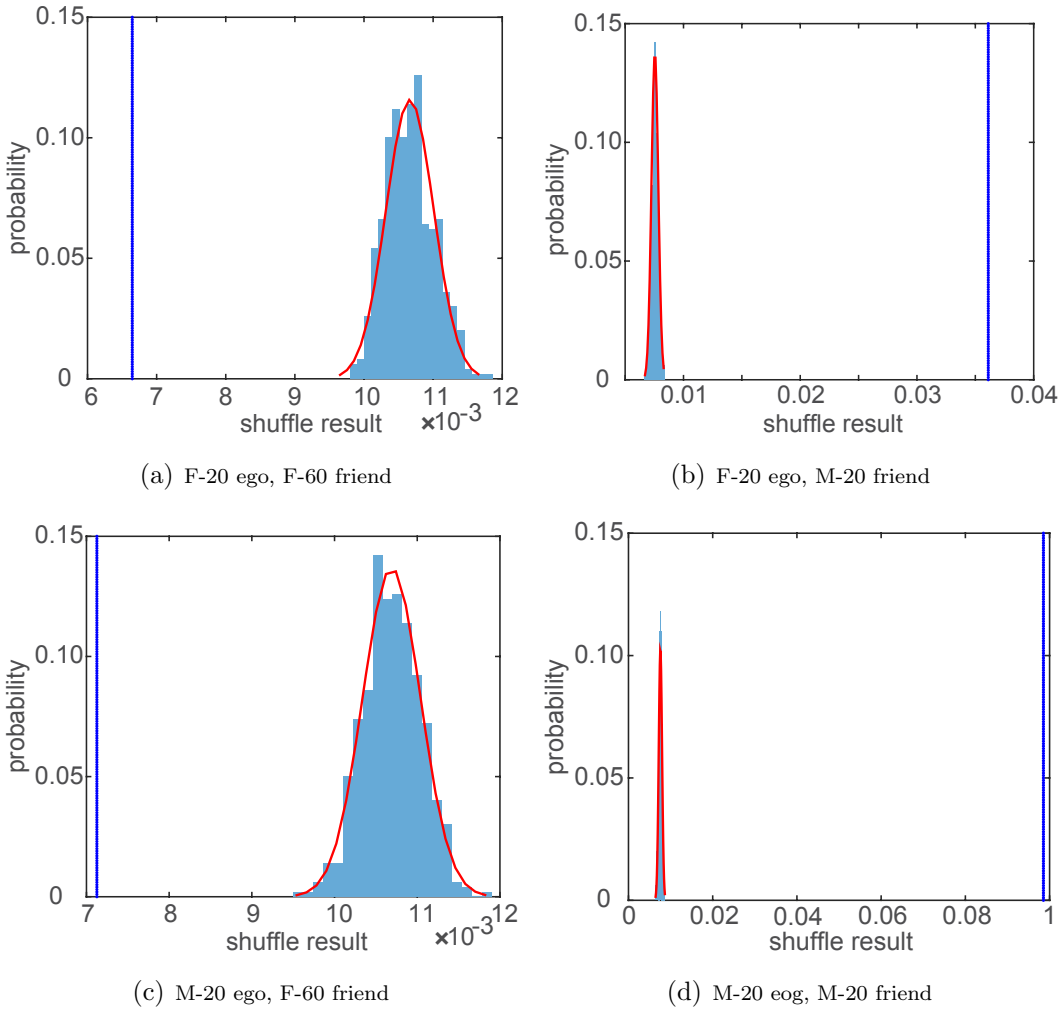
## 2.5 The Null Model in Attributed Networks

We validate the statistical significance of the social strategies observed in the CALL and SMS networks in Section 2.4 by using a null model. The idea of the statistical test is to compare the demographic-based observations $x$ from empirical data to those $\{\tilde{x}\}$ provided by the null model, wherein the demographic profiles of users are randomly shuffled [48, 110]. On the null model, we first randomly assign the demographic profiles of the users on the underlying communication networks, and then observe the social strategies that are derived from the randomly allocated user demographics. We simulate the random process 10,000 times and get the mean $\mu(\tilde{x})$ and standard deviation $\sigma(\tilde{x})$ of the observations $\{\tilde{x}\}$ on the null model. For example, we use four data points selected from Figure 2.3 to illustrate the statistical test, that is, two points ($X$=20, $Y$=60) and ($X$=20, $Y$=-20) from Figure 2.3(a) and 2.3(b), respectively. Figure 2.6 reports the histograms of shuffled results $\{\tilde{x}\}$ of the four points. First, it is clear that the true values $x$ (blue lines) observed from Figure 2.3 largely fall out of the shuffled distributions (histogram plots). Further, we can see that the shuffled distributions are close to the fitted normal distributions (red lines). Accordingly, we use $z$-score to examine the numerical gap between the empirical data $x$ and the randomly shuffled results $\{\tilde{x}\}$ on the null model [193].

$$z(x) = \frac{x - \mu(\tilde{x})}{\sigma(\tilde{x})}$$

A $z$-score of 0 indicates that there exists no difference between empirical data and the null model. A positive (negative) $z$-score represents that the empirical data is over- (under-) represented than expected by chance. In specific, $|z(x)| \geq 3.3$ (corresponding to $p$-value $\leq 0.001$) indicates that the observation from the empirical data is extremely statistically significant.

The statistical tests are conducted for all the social strategies observed on ego net-




(a) F-20 ego, F-60 friend

(b) F-20 ego, M-20 friend

(c) M-20 ego, F-60 friend

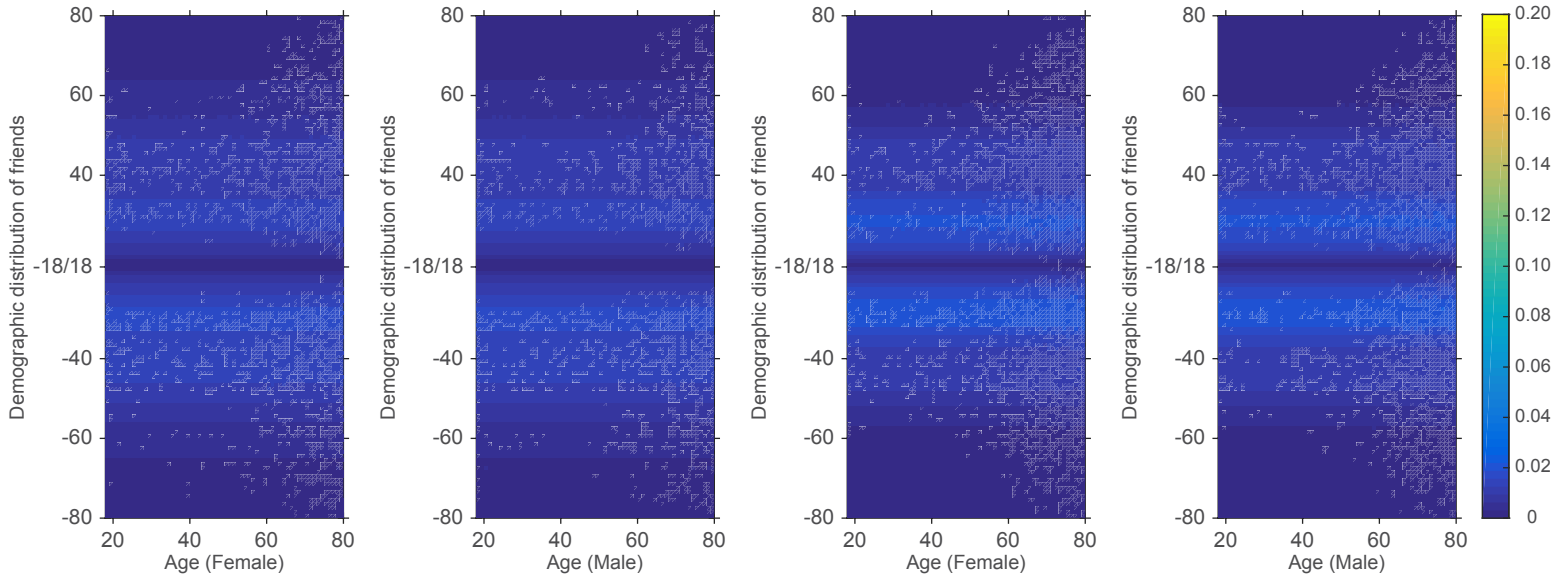
(d) M-20 eog, M-20 friend

Figure 2.6. Illustrative cases of shuffled results and true value in CALL. We select two points from Figure 2.3(a) and two from Figure 2.3(b) to show the shuffled results. Blue line represents the true values from the data (Figure 2.3); blue histograms plot the shuffled results; red line represents the fitted normal density curve.

works, social ties, and social triads in mobile phone call and text messaging behavior. We associate each observation figure of the social strategies presented in Section 2.4 with the shuffled results and $z$-score plots. Specifically, the results on ego networks are shown in Figures 2.7 and 2.8. The shuffled results and $z$-scores on social ties in the CALL and SMS networks can be found in Figures 2.9 and 2.10, respectively. Figures 2.11 and 2.12 present the values of shuffled means and $z$-scores of the social

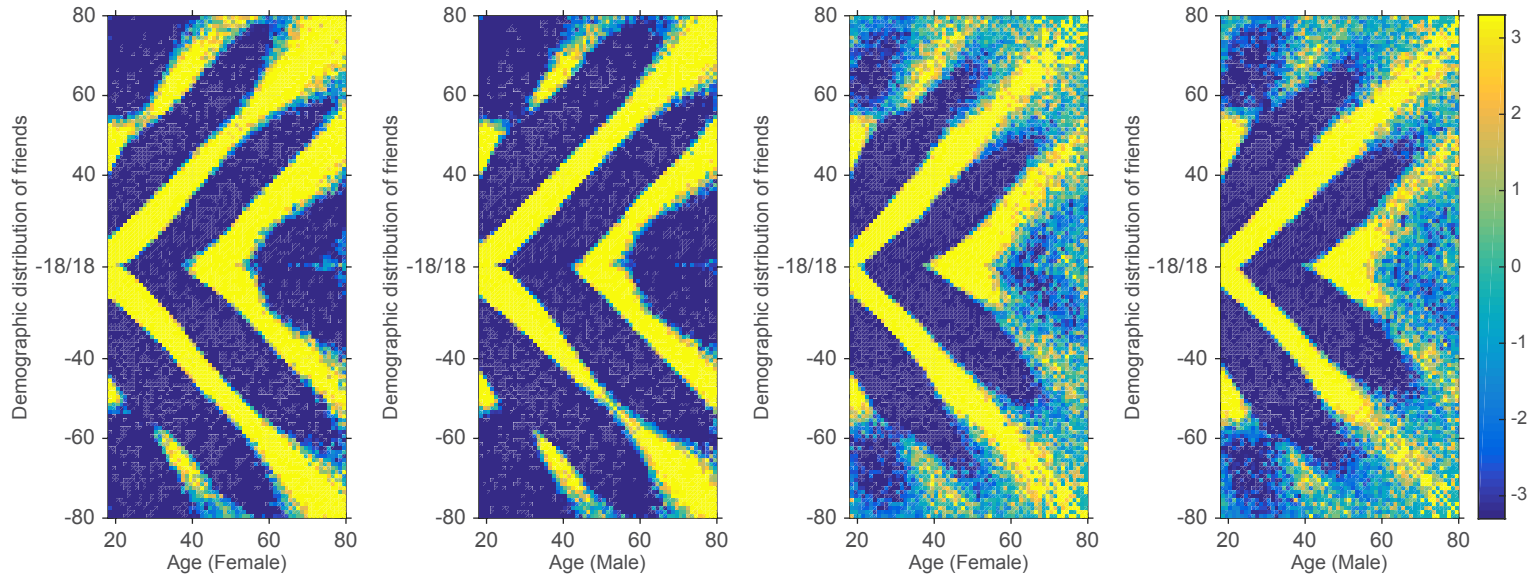
strategies on social triad observed in both the CALL and SMS networks, respectively.

From the figures, we can see that there are large differences between the heatmaps of the observations (data) and those of the means of 10,000 simulating results (shuffle). Moreover, we find that the color of the areas we are interested in from each $z$-score plot tells that $|z(x)| \geq 3.3$. That being said, each social strategy we observed in the mobile communication networks is (extremely) statistically significant.



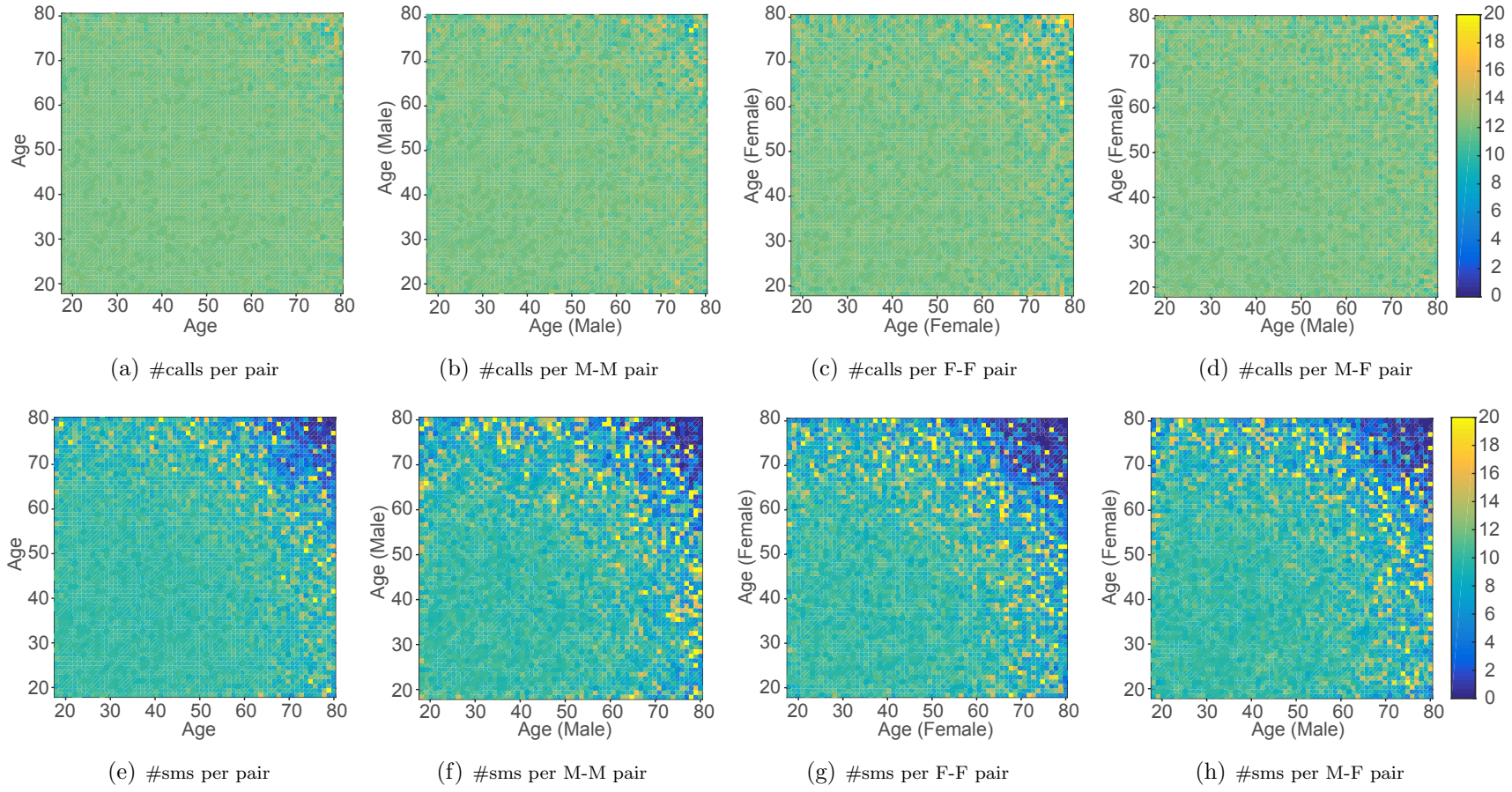

(a) F ego net in CALL (b) M ego net in CALL (c) F ego net in SMS (d) M ego net in SMS

Figure 2.7. Friends' demographic distribution (shuffle). $x$-axis: (a) the age of a female ego in CALL; (b) the age of male ego in CALL; (c) the age of a female ego in SMS; (d) the age of a male ego in SMS. $y$-axis: the age of the ego's friends (positive: female friends, negative: male friends). The spectrum color represents the friends' demographic distribution.




(a) F ego net in CALL (b) M ego net in CALL (c) F ego net in SMS (d) M ego net in SMS

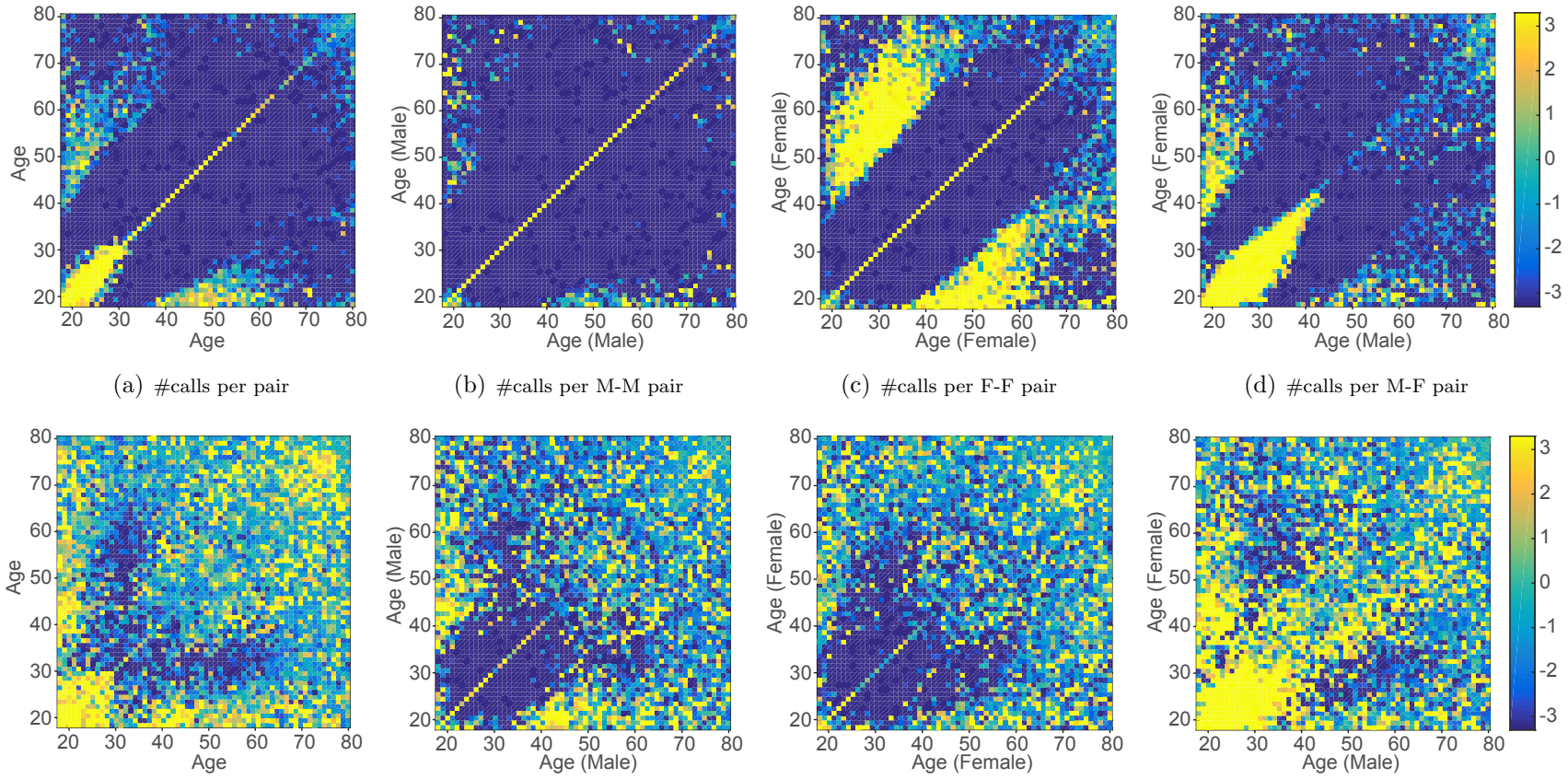
Figure 2.8. Friends' demographic distribution ($z$-score). $x$-axis: (a) the age of a female ego in CALL; (b) the age of male ego in CALL; (c) the age of a female ego in SMS; (d) the age of a male ego in SMS. $y$-axis: the age of the ego's friends (positive: female friends, negative: male friends). The spectrum color represents the friends' demographic distribution.




(a) #calls per pair (b) #calls per M-M pair (c) #calls per F-F pair (d) #calls per M-F pair

(e) #sms per pair (f) #sms per M-M pair (g) #sms per F-F pair (h) #sms per M-F pair

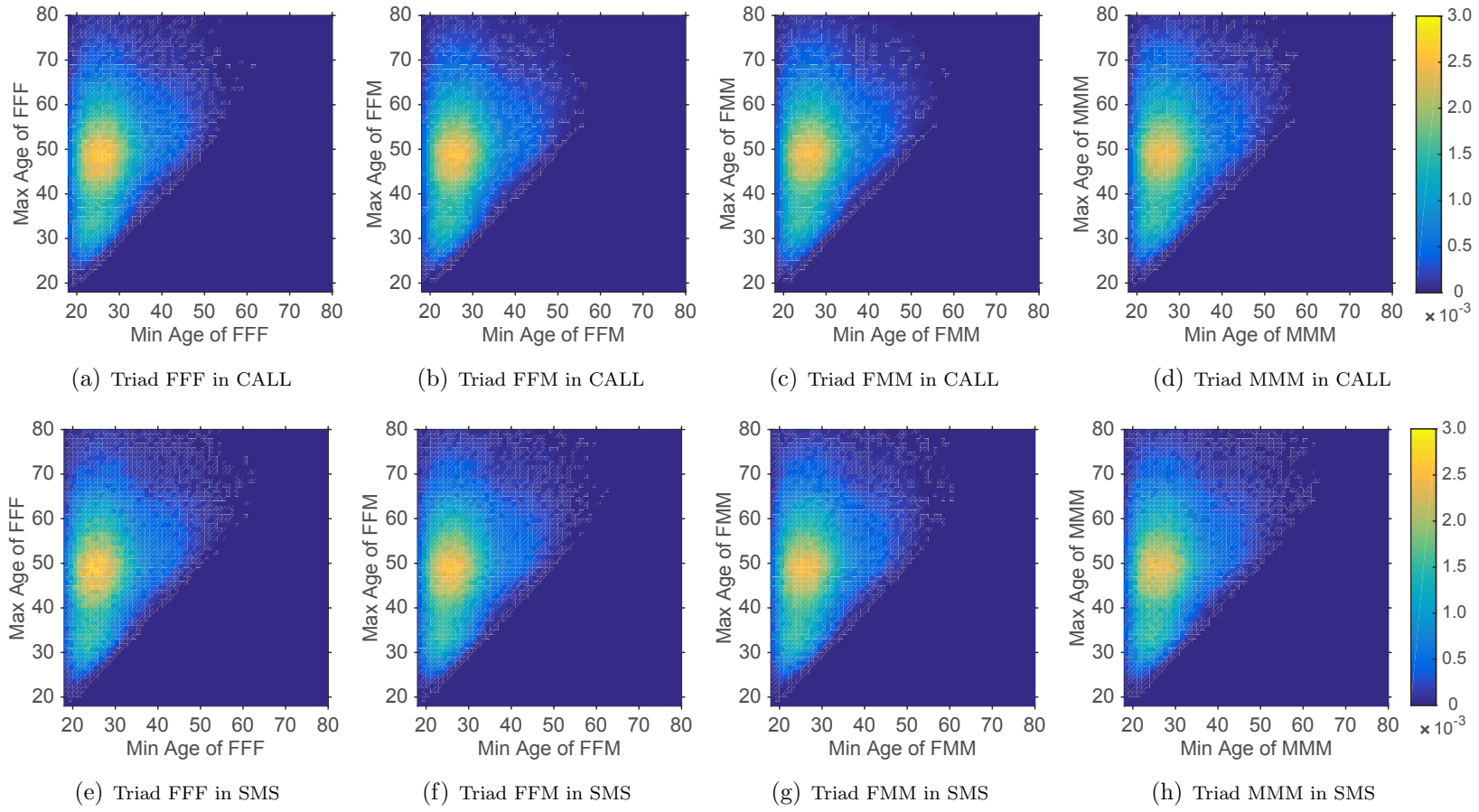
Figure 2.9. Strength of social ties in the CALL and SMS networks (shuffle). $x$- and $y$-axis: age of users with specific gender. The spectrum color represents the number of calls (messages) per month. (a), (b), (c), (e), (f), and (g) are symmetric.




(a) #calls per pair (b) #calls per M-M pair (c) #calls per F-F pair (d) #calls per M-F pair

(e) #sms per pair (f) #sms per M-M pair (g) #sms per F-F pair (h) #sms per M-F pair

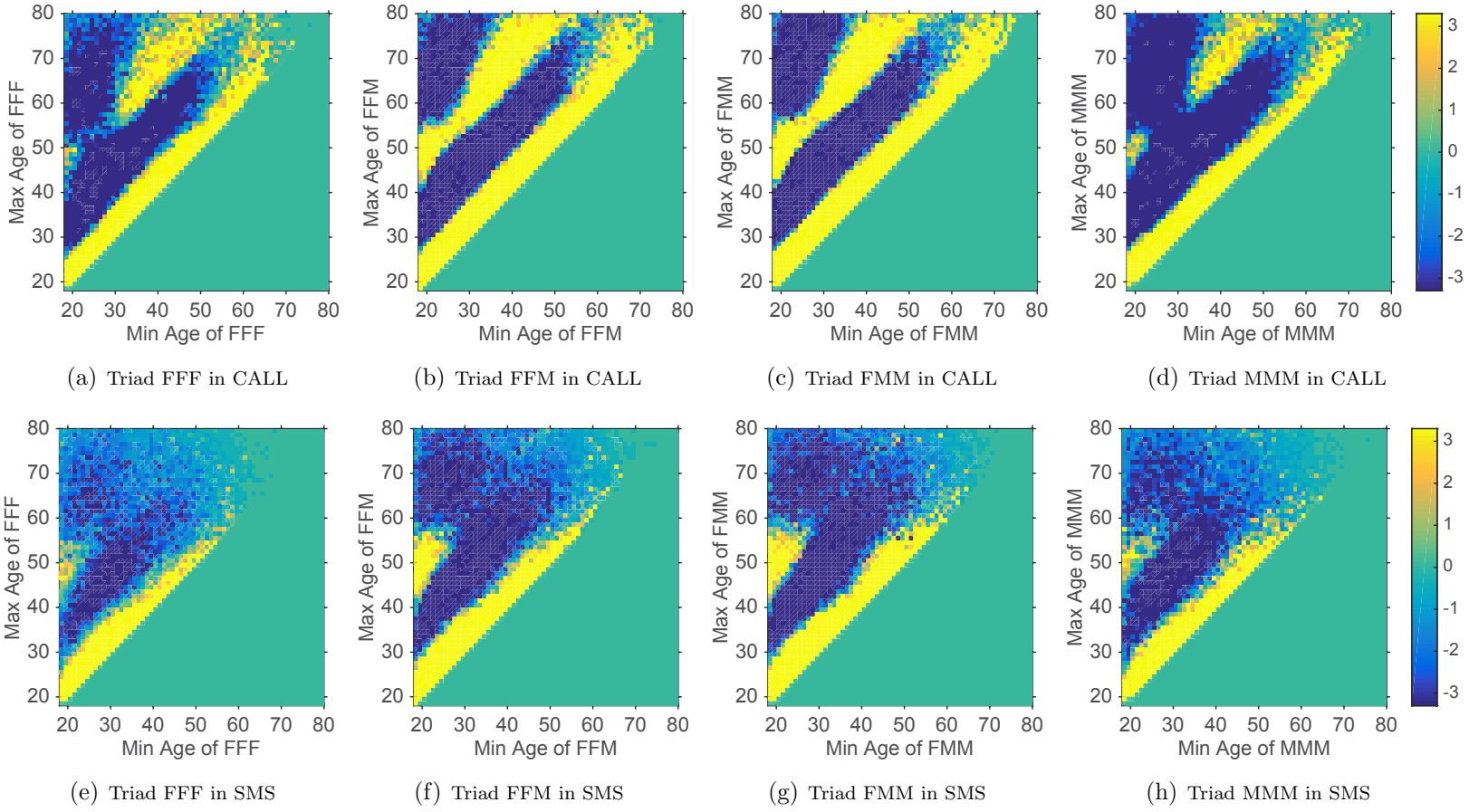
Figure 2.10. Strength of social ties in the CALL and SMS networks ($z$-score). $x$- and $y$-axis: age of users with specific gender. The spectrum color represents the number of calls (messages) per month. (a), (b), (c), (e), (f), and (g) are symmetric.




(a) Triad FFF in CALL (b) Triad FFM in CALL (c) Triad FMM in CALL (d) Triad MMM in CALL

(e) Triad FFF in SMS (f) Triad FFM in SMS (g) Triad FMM in SMS (h) Triad MMM in SMS

Figure 2.11. Social triad distribution in the CALL and SMS networks (shuffle). $x$-axis: minimum age of three users in a triad. $y$-axis: maximum age of three users. The spectrum color represents the distributions.




(a) Triad FFF in CALL (b) Triad FFM in CALL (c) Triad FMM in CALL (d) Triad MMM in CALL

(e) Triad FFF in SMS (f) Triad FFM in SMS (g) Triad FMM in SMS (h) Triad MMM in SMS

Figure 2.12. Social triad distribution in the CALL and SMS networks ($z$-score). $x$-axis: minimum age of three users in a triad. $y$-axis: maximum age of three users. The spectrum color represents the distributions.

## 2.6 Conclusion

According to our comprehensive analysis on the interplay of demographic profiles and mobile communications, we unveil striking gender- and age- based networking differences, which reflect the dynamic social strategies that evolve as a function of the balance between different social needs across lifespans. In summary, we provide the following social phenomena relating to mobile communications:

- Figure 2.2 demonstrates that younger people are active in broadening their social connections, while older people have the tendency to maintain smaller but more closed connections.
- Figure 2.3 confirms demographic homophily, that being said, people tend to interact with others with similar gender and age in both phone call and text messaging channels.
- Figure 2.4 shows that cross-gender social relationships exhibit more frequent communications than the same-gender ones, and the cross-generation interactions are maintained to pass the torch of family, workforce, and human knowledge from generation to generation in social society.
- Figure 2.5 unveils that people tend to expand their social connections with females and males alike during younger and more dating-active period, and put more social investment on maintaining same-gender social groups after entering into middle-age.
- In addition, the gap between the younger and older people in text-messaging channel (e.g., Figure 2.5(e)) is larger than that in phone calls (Figure 2.5(a)), while the difference between males and females (e.g., Figure 2.4(b) vs. 2.4(c)) in phone-call channel are more significant than that in messaging communications (Figures 2.4(f) vs. 2.4(g)).

Despite the promising discoveries of the present work, there is still large room left for future work. First, although we examine the social strategies in two large-scale mobile networks with millions of users, the results are limited to the data we used, that is, the mobile communications from one specific country. Second, there may exist variances on social strategies used by people across different cultural backgrounds, political systems, and geographical boundaries. Therefore, it is natural to

examine the observed results in other countries upon the available data. Third, although previous studies have demonstrated that mobile communications can be used as a proxy to represent human communications, it would generalize our findings beyond mobile channels if online social networks with demographic information could be investigated. Finally, mobile communications are associated with dynamic information, making it necessary to further couple our studies between network structures and user demographics with social dynamics.

# CHAPTER 3

# AGE-SPECIFIC SMALL WORLDS

## 3.1 Overview

In this chapter, we investigate the phenomenon of "age-specific small worlds" using data from the large-scale mobile communication network used in last chapter approximating interaction patterns at a societal scale. Rather than asking whether two random individuals are separated by a small number of links, we ask whether individuals in specific age groups live in a small world in relation to individuals from other age groups. Our analysis shows that there is systematic variation in this age-relative small world effect. Young people live in the "smallest world," being separated from other young people and their parents generation via a smaller number of intermediaries than older individuals. The oldest people live in the "least small world," being separated from their same age peers and their younger counterparts by a larger number of intermediaries. Variation in the small world effect is specific to age as a node attribute (being absent in the case of gender) and is consistently observed under several data robustness checks. The discovery of age-specific small worlds is consistent with well-known social mechanisms affecting the way age interacts with network connectivity and the relative prevalence of kin ties and non-kin ties observed in this network. This social pattern has significant implications for our understanding of generation-specific dynamics of information cascades, diffusion phenomena, and the spread of fads and fashions.

This chapter is largely extracted from a pre-print manuscript [52]. It is a joint work with Omar Lizardo and Nitesh V. Chawla.

## 3.2 Introduction

The fact that any one individual may be capable of reaching any other one via a relatively short chain of network intermediaries is a surprising property of human social networks [227, 228]. This "small world" phenomenon was first documented in a series of classic contact-tracing experiments conducted by Travers and Milgram in the 1960s [147, 212], with a recent large-scale Internet-based replication using a cross-nationally diverse population producing results encouragingly close to those of the original study [41]. More recently, with the increasing availability of large-scale network data built from digitally recorded traces of human communication [55, 117], the existence of the small-world phenomenon has been successfully established using observational data obtained from large-scale systems featuring millions of actors (nodes) and billions of links [12, 92, 120]. One attractive feature of this approach is that it allows for direct calculation of the average number of links separating any two individuals at very close to the whole network level (e.g. the largest connected component in the system). This helps to overcome the key limitation of first generation research on the small world: namely reliance on indirect inference from completed chains obtained from the initial subset of seed nodes. Instead, in large-scale small world research the average of all shortest paths in the network can be calculated directly, although not without computational cost [12, 120].

While useful for demonstrating the robust existence of an important property of social networks, a focus on global estimates of the existence of the small world property has to rely on averages taken over all nodes in the network irrespective of node attributes. The disadvantage of this approach is that it may hide structured heterogeneity in the extent to which different node classes are actually well-represented by the average. This becomes more relevant when we consider that people tend to select contacts with similar social characteristics as themselves [115, 140], a tendency that is reproduced in the sort of electronic telecommunication platforms that have been the

subject of recent attention [21, 44, 110]. Because links are not assigned randomly to node-classes, neither are the number of intermediaries separating a given person from others of the same (or different) class. In this respect, in the context of human social networks, it may be more meaningful to investigate the existence of more targeted realizations of the small world phenomenon, especially with respect to node classes defined by socially significant attributes such as age, gender, and in some contexts, race and social class.

As a first step in this direction, in this chapter we investigate the phenomenon of "age-specific small worlds." Rather than asking whether any two randomly chosen individuals are separated by a small number of intermediaries, we ascertain the extent to which individuals in *the same age group* tend to live in a small world in relation to individuals in the same or other generational clusters. We select age as a focal attribute because it is one of the two (gender being the other one) most powerful traits structuring interaction and sociability in human groups [25, 35, 135, 167].

### 3.2.1 Age, Social Networks, and the Small World

What sort of pattern should we expect to observe in terms of the relative strength of the small world phenomenon across age groups? Sociological research on the connection between the age and kin structure, as well as the relationship between non-kin connectivity and life course transitions can be of help in developing some expectations in this regard. Consider the (idealized) model of the connectivity structure between age and kin groups shown in Fig. 3.1. The figure is meant to encode a series of empirical generalizations taken from relevant work on age, social interaction, and kinship in anthropology and sociology [22, 36, 78, 135, 136, 167]. The basic idea is that the bulk of informal socializing outside of the family occurs within generations following the principle of age homophily [140]. This means that kin ties are the primary link connecting individuals across generations [9, 136]. Kin ties are distinctive because

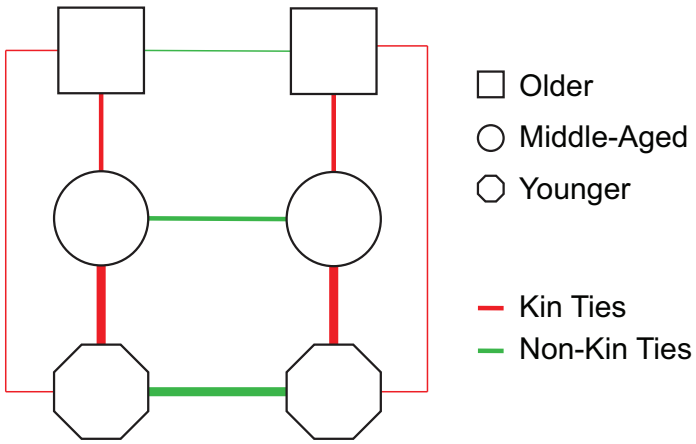



Figure 3.1. Idealized model of the prevalence and strength of kin and non-kin ties across age groups. Shapes represent three generational groups arranged from younger (octagonal), to middle-aged (circle), to older (square). The green edges connecting the shapes represent (idealized) connections among persons who belong to the same age group but who are not biologically related (non-kin ties). The red edges represent (idealized) connections among persons from different age groups who share a biological relation (kin ties). The thickness of the edge indicates the expected relative prevalence and strength (e.g typical communication frequency) for those ties. For the sake of simplicity, cross-generation/non-kin ties are not drawn.

they are largely fixed at birth, are normatively prescribed, and as such display less variation in prevalence and strength across individuals and groups [22, 167]. This has implications for the expected pattern of *cross-generation connectivity* in human societies.

Research in anthropology and sociology points to the historical transformation of the structure of kin ties as societies transition into economic and cultural modernity. As Western (and later non-Western) societies began to industrialize in the the $18^{th}$ and $19^{th}$ centuries, there was a shift towards a "conjugal" (bi-generational) form of family organization [167], and away from tri or quad-generational co-residential living arrangements in which grandparents co-resided with both their children and grandchildren [118]. In this respect, the modal household becomes the bi-generational residence containing only parents and children [178]. In Fig. 3.1, this is indicated

by the thick vertical lines linking the circle (parent) and octagonal (children) generation, and by the relatively thinner vertical links connecting the circle and square (grandparent) and the even thinner lines connecting the octagonal and square.

In addition, note the declining strength of within-generation, cross-kin connectivity as we move up from the youngest to the older groups in Fig. 3.1. This encodes a series of stylized facts from sociological work on the relationship between age and social networks, having implications for the expected pattern of *within-generation connectivity* at the societal level. First, with regards to younger people, sociological work on the subject shows that, free from the demands of work, childcare, and other midlife responsibilities, younger people are better able to devote relatively large amounts of time to within-generation socializing outside the family, increasing their connectivity within this age stratum [134]. In addition, younger individuals tend to spend the majority of their time inhabiting social institutions (such as schools) that encourage same-generation non-kin peer group formation and promote sociable interaction [153]. Second, middle-aged individuals, while continuing to have active dispositions and capacities for socializing with same-age non-kin others, experience a variety of life events that lead to a decline in connectivity. These include transition into marriage, full-time employment, and parenthood [156, 229]. Finally, a long line of research in sociology, anthropology and gerontology demonstrates that older persons experience strong declining attachments to same age peers, with all indicators of sociability experiencing steep drops. These include non-kin contact volume, emotional closeness, and time spent in the presence of others [6, 35, 36, 135, 189, 189]. This also means that as individuals age and lose same-generation non-kin ties, cross-generational connections to children and other relatives come to form a larger proportion of their remaining network [9, 136].

### 3.2.2 Implications for Age-Specific Small Worlds

Because the small world property is premised on the relative connectivity of individuals [8] in relation to others, the existence of combined age and kin effects on social interaction volume should result in predictable consequences for the relative extent to which individuals of different age groups live in a small world. Generally, the less connected the members of a given age group are to others of a given node class (e.g. same or different generation), the less likely they are to be able to reach those others via a small number of intermediaries. Given the empirical patterns encoded in Fig. 3.1, we should then expect that: (*a*) younger individuals should live in the smallest of worlds, especially with respect to same-generation others. In addition, (*b*) given the existence of relatively strong ties to parental generation (via the bi-generational household residence mechanism), they should also be separated by a relatively small (but larger than the same-generation quantity) number of intermediaries from members of the parental generation (and vice versa). However, (*c*) relatively fractured attachments to the grandparent's generation produced by the same bi-generational household structure, should put young people at a longer sociometric distance from their most older counterparts (and vice versa), while (*d*) middle-aged individuals should be in the next "least small" world tier with respect to same-generation peers. That is, their separation from same-generation others should be larger than that of corresponding to their children. Middle-aged individuals, should also (*e*) be relatively close to members of the parental generation via intermediary kin ties. Finally, (*f*) older individuals should live in the "least small" world with respect to same-generation peers, as ties to same-generation others are selectively pruned leaving only kin-tie mediated attachment to middle-aged members of their sons and daughters generation as their primary source of sociability.

## 3.3 Age-Specific Small Worlds

We begin by addressing the question of whether we can identify age-specific small worlds. To do so, we use the large-scale mobile phone data used in Chapter 2 capturing patterns of communication at a societal scale. The data is comprised of more than one billion voice calls and short messaging records spanning two consecutive months—August and September—in the year 2008 representing about one fifth of the population of a large industrialized country. These data are appropriate for our research goals as they have been used profitably in previous studies establishing strong regularities in human communication and mobility behavior [77, 164]. To represent this large-scale communication system as a network, we place an edge between two users if and only if they have reciprocal communications (voice calls or text messages) within the observation time-frame, ensuring that the links capture significant social interactions and relationships.

It is possible that any conclusions regarding age-effects in small-world behavior might be systematically affected by the two month observation window or may not be unique to the generation-specific processes that we outlined earlier. We address these issues in four ways. First, we construct communication networks of increasing temporal scale (using a one-week window), and examine whether our results hold within each cumulative time slice. Second, we also examine whether the values of key quantities, such as the average shortest path lengths, show signs of convergence as we extend the temporal window. Observing such saturation behavior would indicate that two-months are sufficient to extract steady-state properties of the system. Third, we examine whether there are differences in small world behavior across age-by-gender groups, given that gender is a distinct, but equally significant, node-level trait affecting connectivity patterns. Null findings in this respect would provide additional evidence for the generational mechanisms proposed. Finally, we trace patterns of cross-generation connectivity across age-levels and examine whether they provide

evidence for the connectivity mechanisms illustrated the idealized model depicted in Figure 3.1.

### 3.3.1 The Young Live in a Smaller World

The results of the age-specific small world analysis are shown in Figure 3.2 a. We conduct all analyses on the mobile (phone calls and text messages), phone call (CALL), text messaging (SMS) networks, as the results are the same regardless of what communication channel we use as a connectivity criterion (see Figures 3.3 and 3.4). The basic empirical patterns are consistent with expectations. The average shortest path connectivity in the mobile communication network increases steadily with age, until about age 35; it then declines until about age 50 and then rises steadily again into old age. Note that the age markers for the period of increasing "small worldness" for adults (35-50) correspond closely to the ages at which members of advanced industrial societies will be forming "downward" kin ties to their children.

Figure 3.2b shows that age specific average-shortest path distances exhibit the same relative trend with respect to age regardless of the time-window used. To construct the plot, we took the estimates of the average shortest path connectivity of the 50-year-old population as our reference point in each network, putting all time-slice-specific trend-lines in the same scale. The plot shows that the relative small-world gaps between members of different generations is essentially identical across time-windows and can already be observed in the most restricted (one-week) version of the data. These result are consistent with the claim that two-months of communication data is sufficient to establish large-scale regularities with respect to shortest-path behavior in this network. Figure 3.5 provides corroborating evidence for this claim, showing that the accuracy gains of adding additional layers of data decrease dramatically after we cross the three-week cumulative time-slice, with estimates of the network property under investigation (average shortest path) converging around a

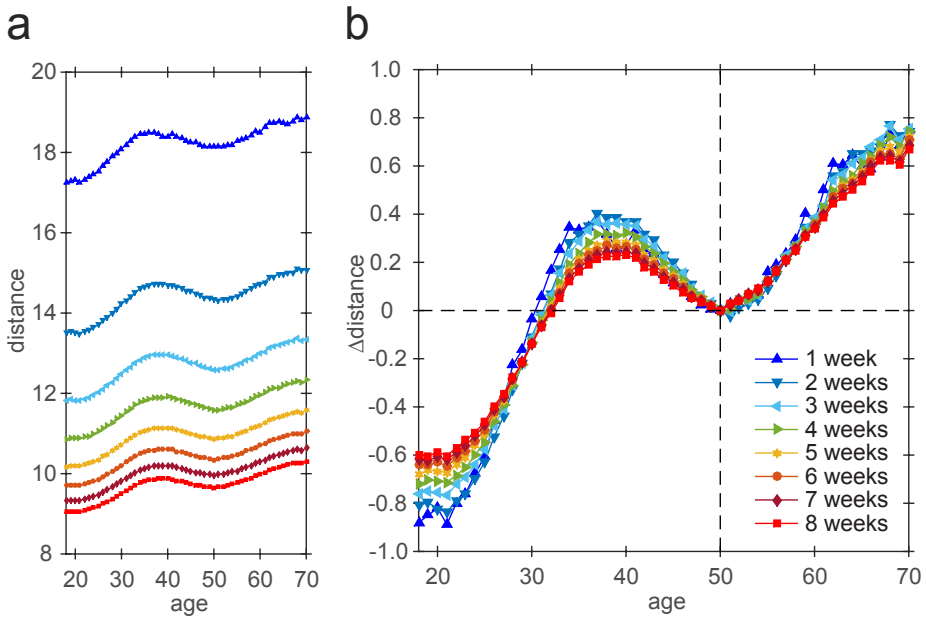



Figure 3.2. Age-specific small worlds across different time-frames in the mobile network. The average degrees of separation vary as a function of age ($a$); The relative variations of age-specific degrees of separation is constant ($b$), that is, in each time-frame the average distance of the 50-year-old people is scaled to 0.

similar steady-state value after the six-week mark (see Figures 3.6 and 3.7 for the results in phone call and text messaging networks).

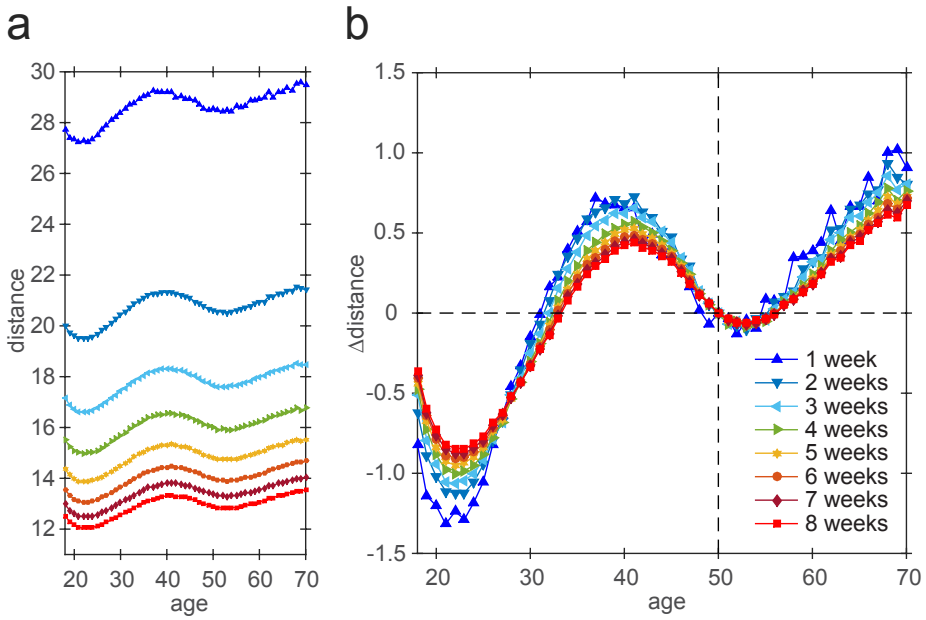



Figure 3.3. Age-specific small worlds across different time-frames in the CALL network.

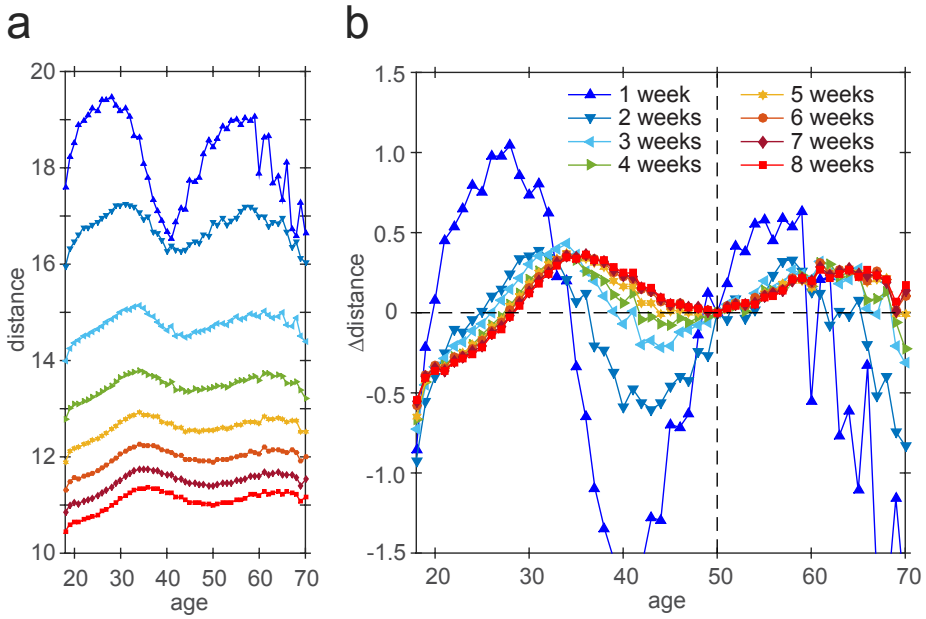



Figure 3.4. Age-specific small worlds across different time-frames in the SMS network.

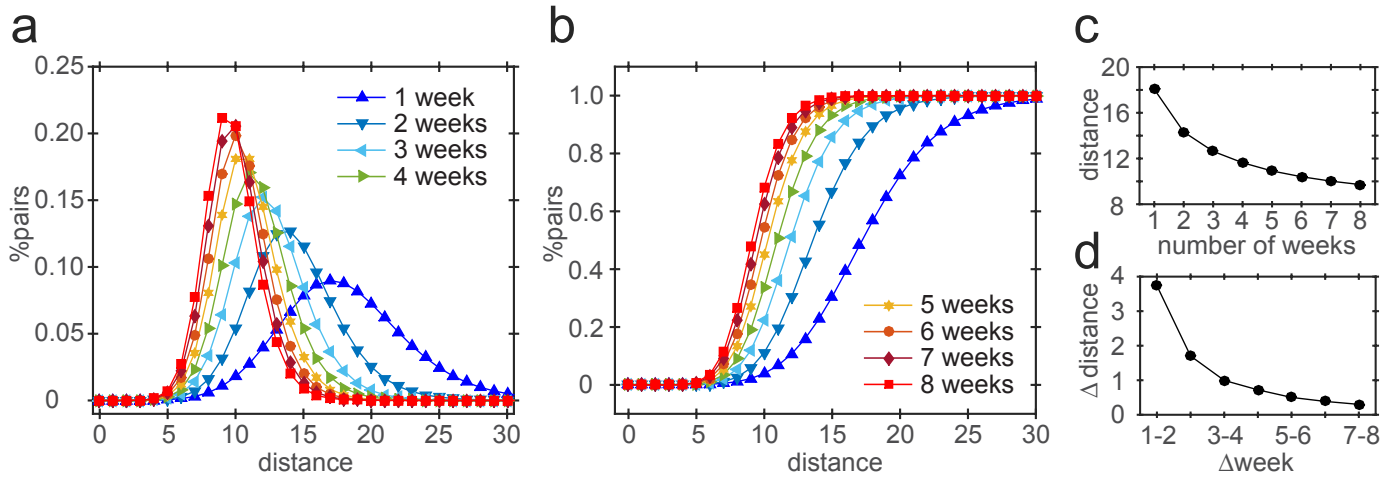



Figure 3.5. Convergence in shortest path estimates with increasing temporal window. The probability mass functions (PMF) of shortest path lengths (distances) across different number of weeks (a); The cumulative distribution functions (CDF) of distances across different number of weeks ($b$); The average distance between each pair of users and time ($c$); The gap between the distances of two consecutive weeks ($d$). For example, when $x$ is 3-4, the corresponding $y$ value represents the difference between the average shortest path lengths observed from the 3-week network and 4-week network.

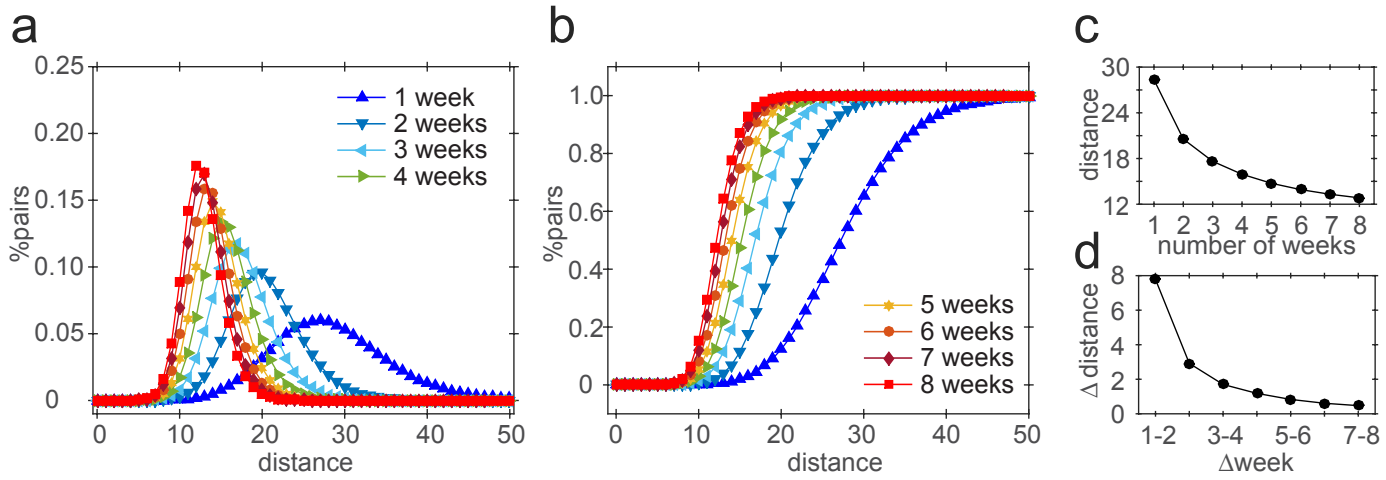



Figure 3.6. Convergence in shortest path estimates with increasing temporal window in the CALL network. The probability mass functions (PMF) of shortest path lengths (distances) across different number of weeks (a); The cumulative distribution functions (CDF) of distances across different number of weeks ($b$); The average distance between each pair of users and time ($c$); The gap between the distances of two consecutive weeks ($d$). For example, when $x$ is $4-3$, the corresponding $y$ value represents the difference between the average shortest path lengths observed from the 3-week network and 4-week network.

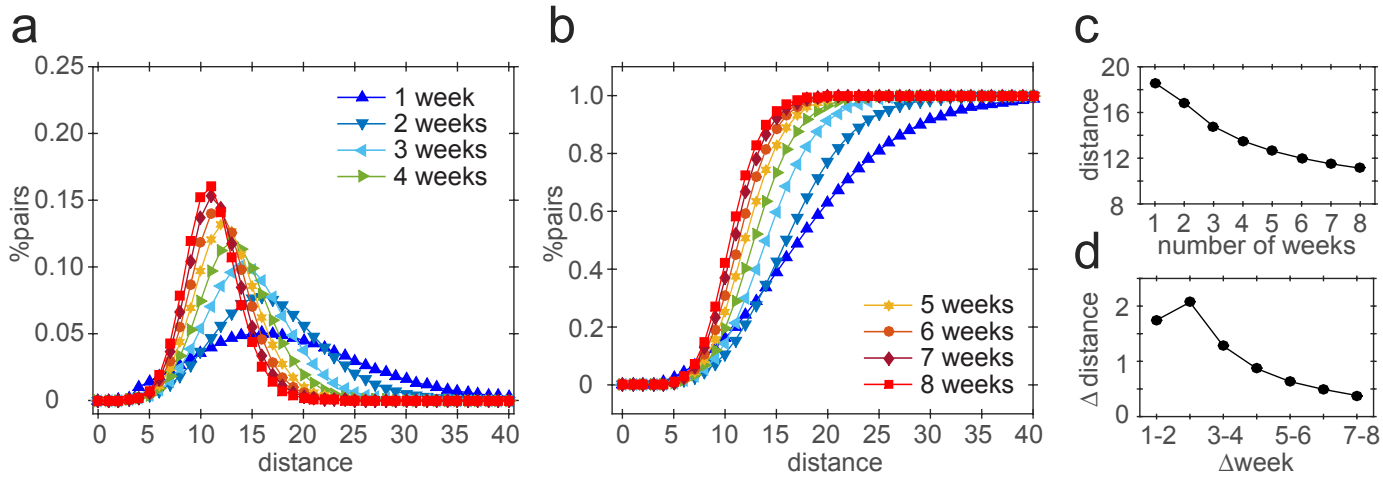



Figure 3.7. Convergence in shortest path estimates with increasing temporal window in the SMS network. The probability mass functions (PMF) of shortest path lengths (distances) across different number of weeks (a); The cumulative distribution functions (CDF) of distances across different number of weeks ($b$); The average distance between each pair of users and time ($c$); The gap between the distances of two consecutive weeks ($d$). For example, when $x$ is $4-3$, the corresponding $y$ value represents the difference between the average shortest path lengths observed from the 3-week network and 4-week network.

### 3.3.2 The Young Are Close to the Young

What are the sources of the small world advantage of young people? To answer this question, we compute average shortest path distances across dyad classes composed of people of different age groups (ranging from 18 through 70). This is shown in Figure 3.8 a. As expected, the small world advantage of younger individuals comes from their relative closeness to their same age counterparts (blue shaded area in the lower left-hand corner of each subplot) coupled with their relative closeness to individuals in their parent's generation (about 20 to 30 years older). This is consistent with sociological work suggesting that the first pattern is due to the formation of non-kin same generation ties (although ties to siblings in the same generation are also included here), while the latter are due primarily to kin ties to parents (and indirectly to other members of the parental generation). Individuals between the ages of 35 and 50 end up being sociometrically closer to their younger counterparts (offspring generation) than they are to their own generation, thus explaining the relative decline in average shortest path distances for individuals within this age range. This result is consistent with sociological research pointing to the disruption of same-generation non-kin ties with middle-aged life transitions, and the relative stability and durability of kin ties to offspring given their non-elective status [167, 229].

As shown in the red-shaded area in the upper-right hand corner of the plot, the reason why older individuals live in the "least small world." is due to their relatively large sociometric distance from members of the same-generation and that of their immediately preceding (offspring) age group. This is consistent with work showing steady decline in sociability and connectivity with in elective (non-kin) ties leaving older persons with non-elective (kin) ties as their only source of connectivity [136]. As with our previous results, relative age-based patterns of cross-generational connectivity observed in the 8-week network are also consistently observed in the 1- to 7-week networks (Figures 3.8d – j). This robustness check shows that the age-specific

small world effect is independent on restrictions on the temporal window covered by our data.

Figure 3.8b shows a heatmap illustrating what happens when we shuffle the demographic attributes of each vertex in the network (leaving both the network structure and the proportion of vertices belonging to a given age group intact) while computing the average shortest path distances across age groups for fifty different realizations of the reshuffled network (see Materials and Methods for details). As shown by the homogeneous coloring across the figure, age-group differences in average shortest-path distances to members of other age groups disappear, and all age groups converge to the average geodesic distance for all pairs in the mobile network ($\mathcal{L} \approx 9.7$). This suggests that, consistent with our account, differences across age groups in "small worldness" emerge as a result of systematic preferences and constraints generating specific within and cross-generation social attachments in human populations [140].

Figure 3.8c shows a heatmap of the distribution of $z$-scores obtained from comparing the observed average geodesic distances across age groups against what we would have expected by chance (from the fifty reshuffled realizations of the network as given in Figure 3.8b). The results confirm that younger individuals live in smaller worlds in relation to same generation peers and older generation contacts than we would expect by chance, while older individuals live in larger than expected small world in relation to same generation peers and members of the immediately preceding generation (middle-aged individuals).

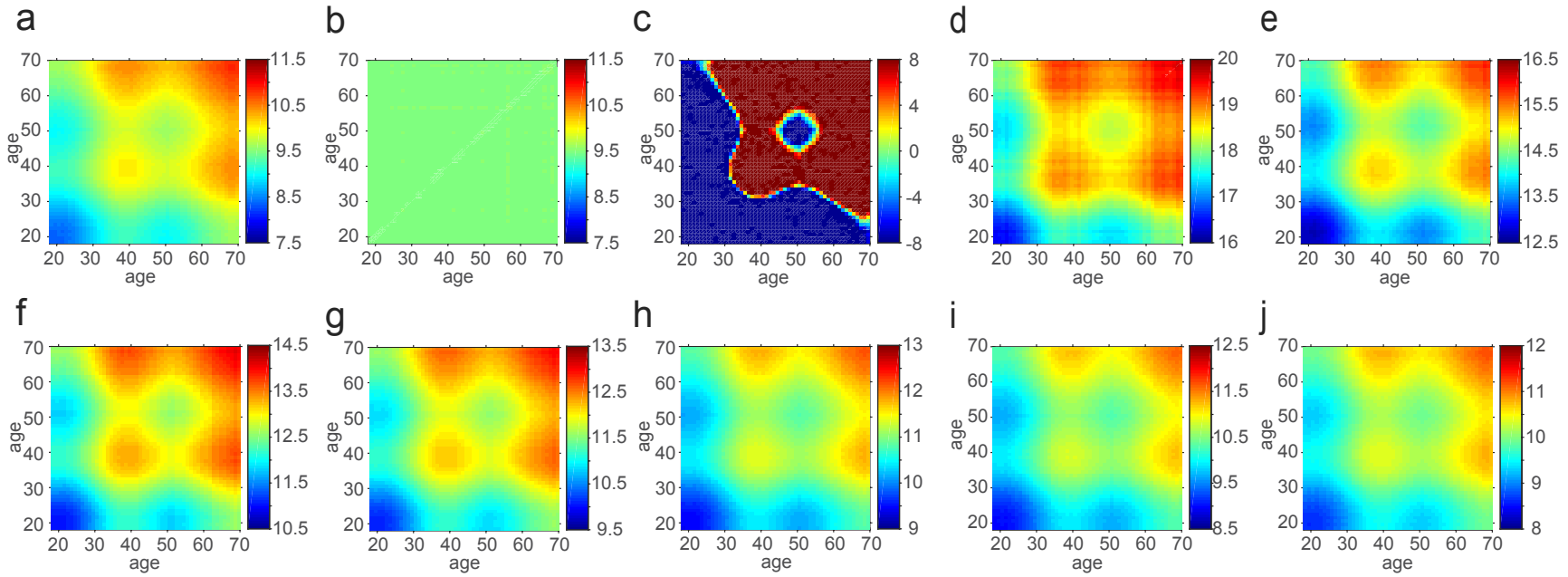



Figure 3.8. Average degrees of separation across age groups. The spectrum color represents the average shortest path length in the 8-week Mobile network ($a$), shuffled average shortest path length ($b$), and $z$-score value ($c$) between two people of age indicated by $x$- and $y$- axes. The spectrum color in figures $d$, $e$, $f$, $g$, $h$, $i$, and $j$ represents the average shortest path length in the 1-, 2-, 3-, 4-, 5-, 6-, and 7-week mobile networks.

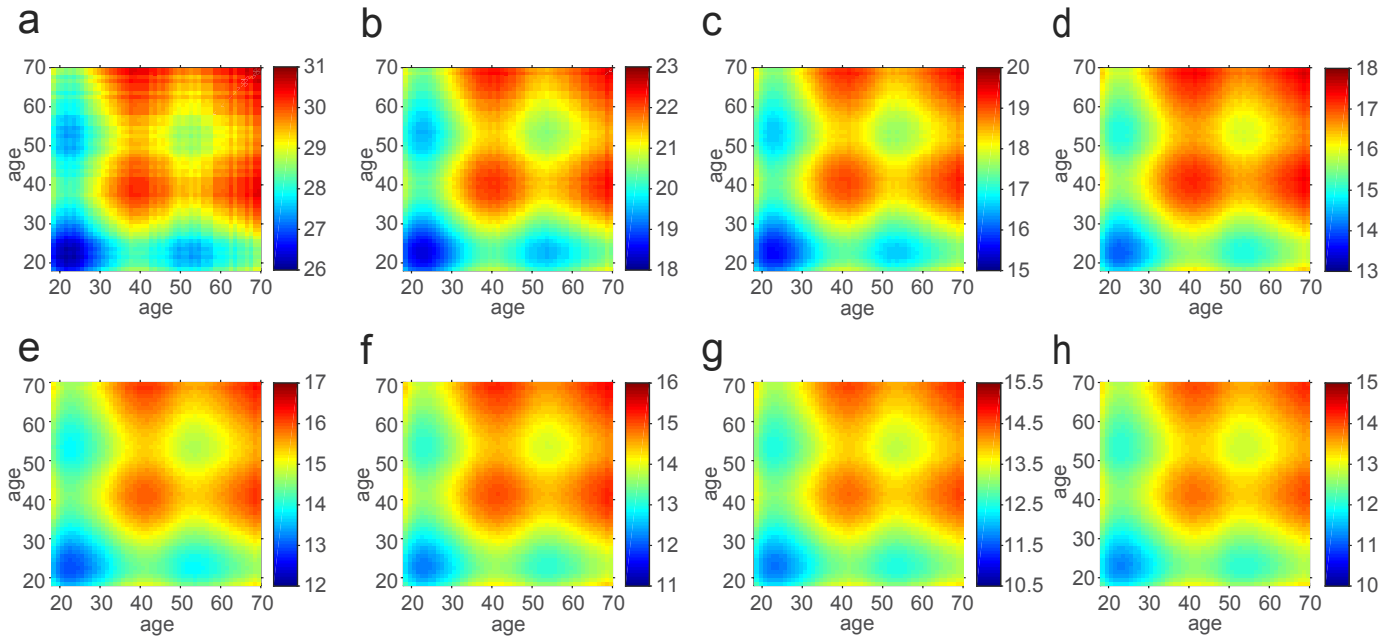



Figure 3.9. Average degrees of separation across age groups in the CALL network. The spectrum color represents the average shortest path length in the 8-week Mobile network ($a$), shuffled average shortest path length ($b$), and $z$-score value ($c$) between two people of age indicated by $x$- and $y$- axes. The spectrum color in figures $d$, $e$, $f$, $g$, $h$, $i$, and $j$ represents the average shortest path length in the 1-, 2-, 3-, 4-, 5-, 6-, and 7-week Mobile networks.

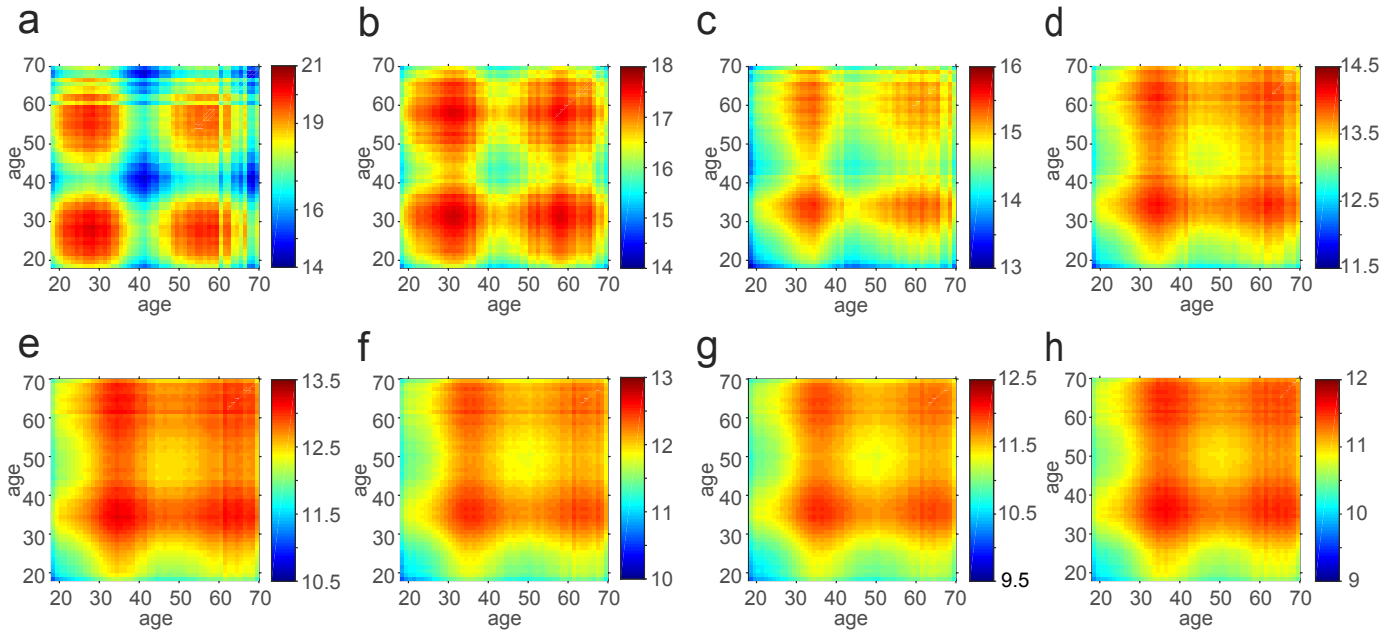



Figure 3.10. Average degrees of separation across age groups in the SMS network. The spectrum color represents the average shortest path length in the 8-week Mobile network ($a$), shuffled average shortest path length ($b$), and $z$-score value ($c$) between two people of age indicated by $x$- and $y$- axes. The spectrum color in figures $d$, $e$, $f$, $g$, $h$, $i$, and $j$ represents the average shortest path length in the 1-, 2-, 3-, 4-, 5-, 6-, and 7-week Mobile networks.

### 3.3.3 Null Gender-Specific Small Worlds

To provide corroborating evidence that the mechanisms generating differences in small world behavior are unique to generational node classes, we investigate whether there are gender effects in age-specific small worlds. Looking at differences between men and women is relevant, since other than age, gender is the one characteristic that has been shown to systematically impact aspects of communication behavior in social networks [110, 156]. However, if the model presented in Figure 3.1 is on the right track we should find little or no evidence of gender-by-age specificity in small worlds, since sociological work shows that the mechanisms generating the age-specific small-world behavior are common to both men and women.

As Figure 3.11 shows, we find that the pattern of decreasing "small-worldness" as persons age is common to men and women (see Figures 3.12 and 3.13 for the results in phone call and text messaging networks). The one exception is the slightly stronger increase in "small worldness" for women between the ages of 30-50 in relation to men of the same age. This pattern of results is consistent with the downward (offspring generation) kin-based connectivity mechanism proposed to explain this effect, as mothers are more likely to maintain regular interactions with their children than fathers. Notably, we find that average shortest path estimates do not differ across dyad pairs classified according to the gender mix (Figures 3.11 b – c). Our analyses show that the shortest path connectivity between two females (F-F), one male and one female (M-F), and two males (M-M) follows highly overlapping distributions (Figure 3.11b) and are not sensitive to time-window restrictions in the data (Figure 3.11c). Figures 3.11d – f show that the relative sociometric distance between nodes based on age-classes does not depend on the gender mix as the same pattern of the young being close to the young being close to their same age peers and the old being far from other old people is replicated for same gender (d, f) and different gender (e) dyad classes. Given that previous studies have shown that cross-

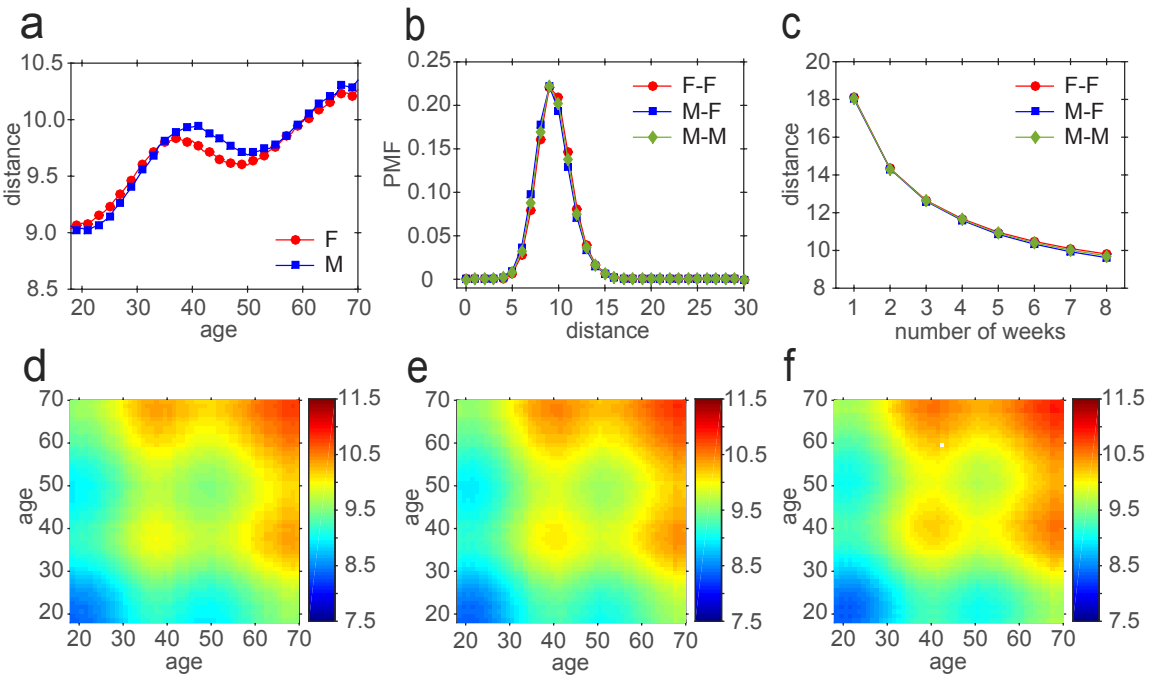



Figure 3.11. Gender-specific small worlds across age groups. The average distances by age do not vary a lot for female (F) and male (M) in the 8-week mobile network ($a$); the probability mass functions of shortest path distances between three different gender pairs overlap with each other in the 8-week network ($b$); The average distances between different gender pairs are the same in all eight networks of different length of time-frames ($c$); the spectrum color represents the average shortest path lengths between two females ($d$), one male and one female ($e$), and two males ($f$) in the 8-week mobile network. The strong similarities among the three heatmaps suggest relative age-specificity of mobile small worlds does not depend on gender in a strong way.

gender interactions are consistently more intense and frequent than those between same-gender pairs in different communication channels [44, 120], our findings suggests that relative small world differences between age groups are not generated by heterogeneity in the characteristic link strengths across age-classes.

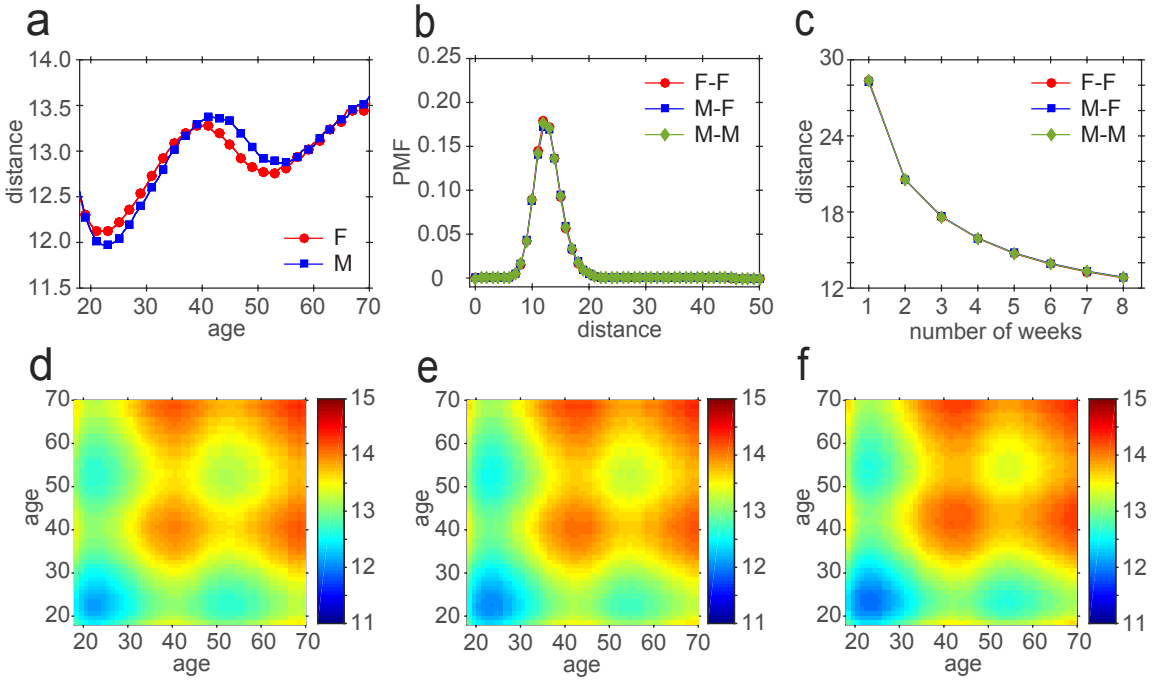



Figure 3.12. Gender-specific small worlds across age groups in the CALL network.

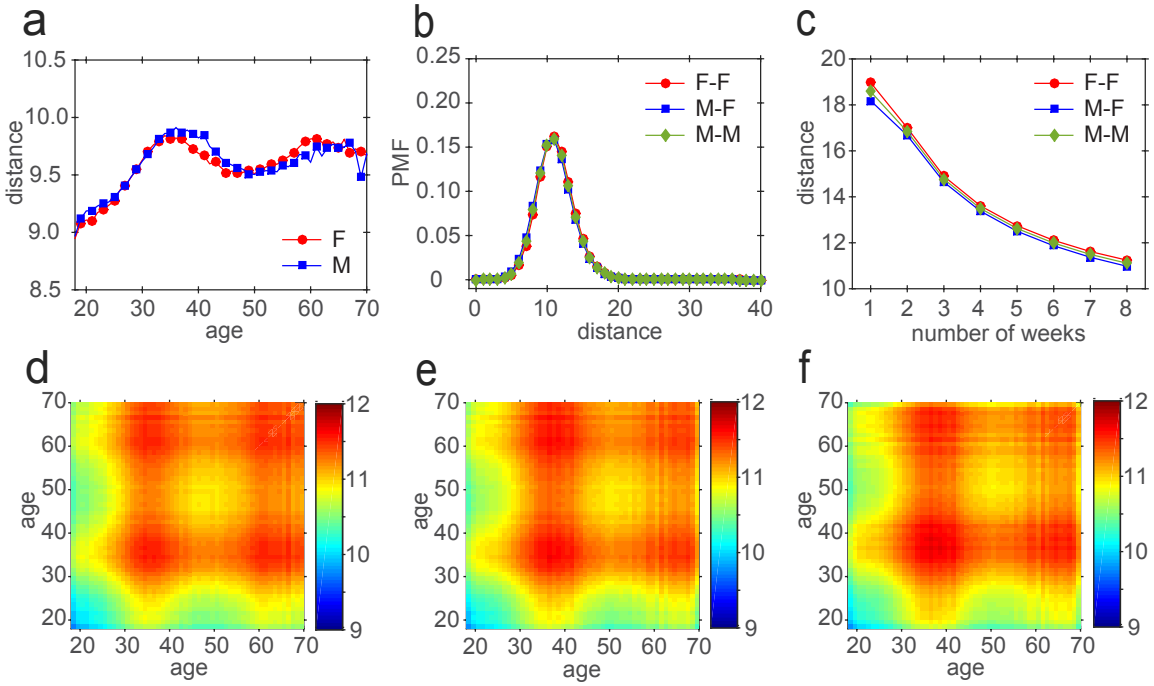



Figure 3.13. Gender-specific small worlds across age groups in the SMS network.

### 3.3.4 Evidence for Proposed Connectivity Mechanisms

As we noted earlier, sociological and anthropological work on age and social networks suggests that the mechanism generating age-specific small worlds is that the cross-age-group connectivity distribution is systematically different for older and younger persons. More specifically, same-generation (primarily non-kin) sociability should steadily decline and be replaced by increasing cross-generation (primarily kin-based) sociability. To examine whether we can observe evidence of this mechanisms in this network, Figure 3.14 a plots the proportion of ages for each age group that link them to same generation (plus or minus five years difference), older generation (between 20 to 30 years older) and younger generation (between 20 to 30 years younger) groups. All rates are calculated from the mobile electronic communication network.

The findings provide strong evidence for the idealized pattern depicted in Figure 3.1, suggesting that these are the mechanisms behind the age-specific small world effects that we observe. Younger individuals (e.g. between the ages of 20 and 35) have relatively high rates of communicative interaction with both their same age peers and those in the immediately preceding (parental) generation. However, as we move up along the $x$-axis, we see a steady decline in same-generation sociability and its gradual replacement by cross-generation sociability (20 to 30 years younger). This is indicative of attrition in same-generation ties for older individuals and their replacement with cross-generation ties to the immediate kin (child) generation. The two lines cross at about 60 years of age, which is close to the institutionally mandated time for retirement from work activity in industrialized Western societies (such as the one from which the mobile network originated), providing further support for the model.

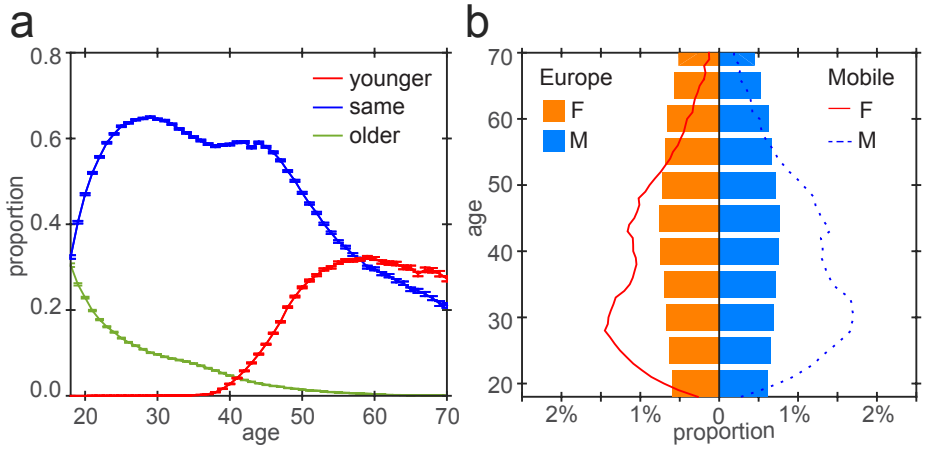



Figure 3.14. Connectivity mechanisms behind age-specific small worlds. The proportion of one's contacts of different age groups conditioned as a function of the person's own age ($a$). Specifically, one's contacts of the "same" generation are denoted as those aged between $x$-5 and $x$+5, where $x$ represents his or her age, the "older" generation aged between $x$+20 and $x$+30, and the younger generation aged between $x$-30 and $x$-20 (The mean values are observed at a 95% confidence interval); The population distribution observed from the mobile data is different from the European population distribution at the same year, that is, 2008.

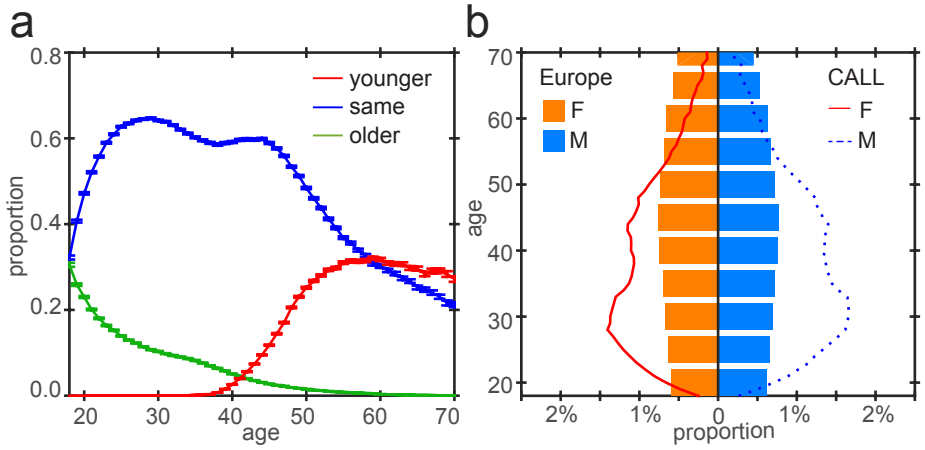



Figure 3.15. Connectivity mechanisms behind age-specific small worlds in the CALL network.

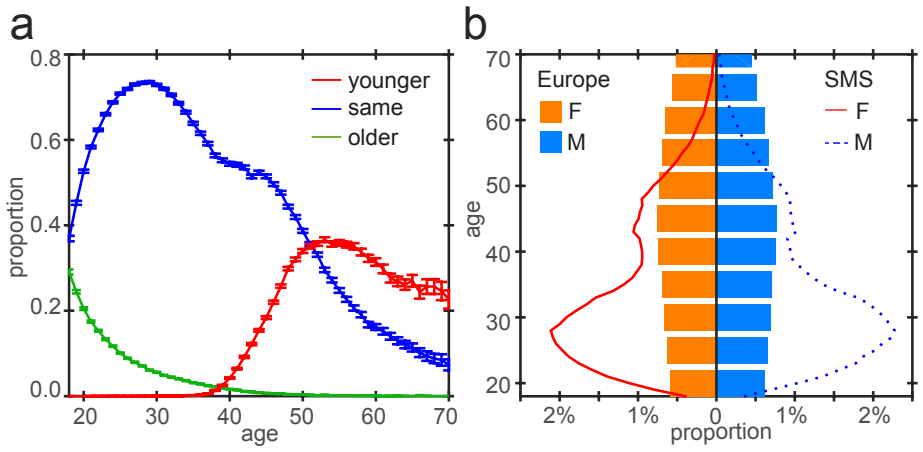



Figure 3.16. Connectivity mechanisms behind age-specific small worlds in the SMS network.

## 3.4 Materials and Methods

### 3.4.1 Mobile Phone Networks

We use the mobile phone data used in the previous chapter, which spans between August and September in 2008 [44, 58, 77]. To investigate the evolution of mobile small worlds, we choose to use one week as the time unit to create networks of different durations (weeks). Specifically, we use the first week (Monday, August $4^{th}$ to Sunday, August $10^{th}$, 2008) of communication logs to construct the 1-week network. Similarly, the $k$-week network ($1< k \leq 8$) was extracted from the first $k$ consecutive weeks of data, meaning that the first day of each network always starts from Monday, August $4^{th}$, 2008 and the last day of the 8-week network ends at Sunday, September $28^{th}$, 2008. Further we extract the giant component as the experimental network from each network [12, 120, 164]. In this way we construct eight mobile networks of different length of durations from the communication logs, with the largest and longest-spanning network, the 8-week one composing of 5,171,066 nodes and 9,885,493 undirected edges. The order and size of the eight mobile, phone call, and text messaging networks are listed in Table 3.1. Our visualization shows that

the combined mobile networks obey the densification power law [121] with a good fit, that is, the number of edges grow superlinearly in the number of nodes in mobile communication networks (see Supplementary Figure 3.17).

In the resulting 8-week mobile network, 89% of nodes are associated with the corresponding users' gender and age information. We calculate the shortest paths between all pairs of users and report the results between those with known gender and age attributes.

### 3.4.2 Shortest Paths in Big Networks

In this work, rather than employing the sampling and probabilistic methods used in previous work [12, 120], we instead leverage a 48-core CPU computing server to determine the *exact* shortest path length between *all* pairs of users, that is $s = n \times (n-1)/2$ pairs, where $n$ is the number of nodes in each network. We use the parallel breadth-first search algorithm to compute the shortest paths between all pairs of users, and more essentially, during each step of search, to record the length of the shortest path between two users specified by their gender and age information. In the parallel algorithm, $n/48$ nodes' distances to all $n$ nodes are allocated to one CPU for computation. For example, to compute the shortest path distances between $s \approx 1.33 \times 10^{13} (n = 5,171,066)$ pairs of users in the largest mobile network (8-week), each CPU is responsible for $s/48 \approx 2.8 \times 10^{11}$ pairs of users. By using the computer server with Quad 12 core 2.3 GHz Intel Xeon CPUs E7-4850 (48 cores in total), we are able to compute the exact shortest path length between all pairs of users within 37 hours for the 8-week mobile network.

TABLE 3.1

THE STATISTICS OF EIGHT MOBILE NETWORKS

| | #weeks | 1-week | 2-week | 3-week | 4-week | 5-week | 6-week | 7-week | 8-week |
|---|---|---|---|---|---|---|---|---|---|
| MOBILE | #nodes | 1,406,743 | 2,698,575 | 3,444,931 | 3,958,354 | 4,371,045 | 4,686,770 | 4,948,254 | 5,171,066 |
| | #edges | 1,672,693 | 3,659,144 | 5,119,451 | 6,314,822 | 7,402,307 | 8,336,223 | 9,153,808 | 9,885,493 |
| CALL | #nodes | 683,422 | 1,856,733 | 2,587,069 | 3,087,363 | 3,479,397 | 3,771,458 | 3,996,406 | 4,176,011 |
| | #edges | 786,952 | 2,385,335 | 3,606,013 | 4,600,727 | 5,499,004 | 6,266,175 | 6,915,723 | 7,482,933 |
| SMS | #nodes | 159,745 | 570,219 | 972,996 | 1,305,151 | 1,596,555 | 1,838,026 | 2,052,003 | 2,241,307 |
| | #edges | 172,657 | 639,635 | 1,141,875 | 1,596,942 | 2,026,486 | 2,411,111 | 2,770,033 | 3,101,637 |

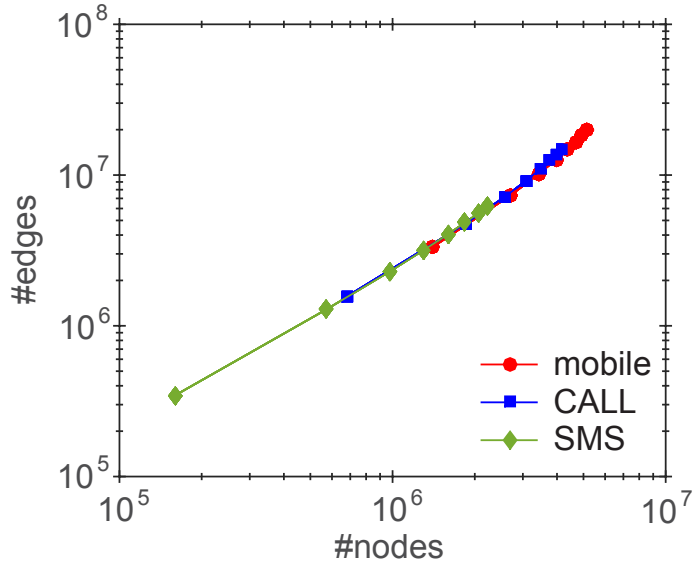



Figure 3.17. The size of mobile phone networks as a function of their orders in log-log scales. Three networks extracted from different channels obey the densification power law [121] with a close slopes.

## 3.5 Discussion and Conclusion

Previous work has shown the "small world" property to be a counter-intuitive but robust signature of human social networks. The classic work by Milgram [147] as well as more recent replications using email chains [41] used experimental strategies aimed at inferring average network diameter from the average length of completed chains. More recent work using social media and electronic communication data allows for the computation of average shortest paths at a societal or even "planetary" scale [12, 92, 120]. However, most of this work remains focused on the small world property as a feature of the entire network, but has not looked at vertex-level heterogeneity in the existence of this property.

In this chapter, we ask the question of whether there are age-specific small worlds. Using a large-scale mobile communication network that approximates the volume of communication of a large-scale industrialized society, we ask not whether any two random individuals are separated by a relatively small number of intermediaries.

Rather, we ask whether individuals in different age groups live in more or less small worlds in relation to members of other groups. This question is important since age is one of the few characteristics that has been shown to structure human interaction in all human societies [157].

Our results reveal systematic heterogeneity in the extent to which people of different ages can be said to live in "more or less" small worlds. The pattern of this heterogeneity is, in its turn, predictable from well-known regularities uncovered in anthropology and sociology related to the relationship between age, sociability, changing structure of generational living arrangements in modern societies, and the relative rates of kin-based and non-kin-based connectivity throughout the life-course (see Figure 3.1). Younger individuals live in the smallest of worlds, both in relation to same-age peers and cross-generation consociates, while older individuals live in the "least small" of worlds, being particularly likely to be separated by a larger number of intermediaries from same-generation peers.

These results have important implications because the small world property of human social networks lies behind a variety of phenomena associated with processes of cultural transmission, the emergence of information cascades, and other diffusion processes. As a rule, shortest connectivity paths between persons facilitate the fast spread of information and thus contribute to the large-scale adoption of novel beliefs, behaviors, practices, and products [154, 227]. Our results thus imply that in any given modern society, due to their greater sociometric proximity to both same-generation and cross-generation others, younger persons are more likely to serve as the most effective seeds and most likely conduits for the rapid spread of novel information, behaviors, practices, and any other element that may be subject to "contagion" and diffusion dynamics. Older individuals, on the other hand, due to their greater sociometric isolation from other age groups, are the least likely to play this role. This implication is consistent with work in sociology and marketing showing that

such phenomena as fads, fashions, and information/behavior cascades occur more frequently among the young [75], and that members of older generations are generally dependent on younger individuals to keep abreast of novel behaviors, products, and activities [74]. Our work thus reveals that these long standing observations have an intuitive basis in the sociometric location of the young in relation to the old.

Our results also imply that greater sociometric isolation of older individuals will result in their being the last to hear or be exposed to novel "viral" practices, beliefs, and objects net of any dispositional "conservatism" that may come with advanced age [38]. Thus even older individuals who may be potentially open to new experiences and be likely candidates for the adoption of innovations, will be at a structural disadvantage. However, our argument and results do suggest that *if* older persons do experience exposure it is more likely to come from cross-generational next of kin ties (most likely children) than from non-kin same-generation peers. In this respect, the existence of various "generation gaps" in attitudes, behaviors, and practices may be as much of a product of the qualitatively distinct social structural position of the young and the old as it is of cohort-based, period-based, or aging-dynamics.

In this chapter, we have provided a model and a set of tools for how to investigate heterogeneity in "generic" properties of large-scale networks across vertex attributes. Future work can build on our current effort and examine the extent to which heterogeneity in the small world (and other well-defined network properties) that have been primarily investigated irrespective of the categorical attributes of vertices in human social networks do vary in a structured way according to those attributes, while outlining the implications of this variation for our understanding of important structural and dynamics processes in such networks.

# CHAPTER 4

# DEMOGRAPHIC PREDICTION IN NETWORKS

## 4.1 Overview

Demographics are widely used in marketing to characterize different types of customers. In previous two chapters, we discover the correlations between user demographics and network structures. In this chapter, we further study to what extent users' demographic profiles can be inferred from their mobile communication patterns. Specifically, we formalize the demographic prediction problem of inferring users' gender and age simultaneously. We propose a factor graph-based *WhoAmI* method to address the problem by leveraging not only the correlations between network features and users' gender/age, but also the interrelations between gender and age. In addition, we identify a new problem—coupled network demographic prediction across multiple mobile operators—and present a coupled variant of the *WhoAmI* method to address its unique challenges. Our extensive experiments demonstrate both the effectiveness, scalability, and applicability of the *WhoAmI* methods. Finally, our study finds a greater than 80% potential predictability for inferring users' gender from phone call behavior and 73% for users' age from text messaging interactions.

This chapter is largely extracted from previous publications [49, 53]. It is a joint work with Jie Tang, Yang Yang (THU), Jing Zhang, and Nitesh V. Chawla.

## 4.2 Introduction

In this chapter, we study to what extent users' demographic information can be inferred by mobile social networks. We formally define a double-label classification problem. The objective is to simultaneously infer users' gender and age by leveraging their interrelations. This problem is different from traditional classification problems, where only the correlations between the dependent variable $Y$ and feature vector $\mathbf{X}$ are considered. In this problem, we are given two dependent variables $Y$ (gender) and $Z$ (age), and a feature vector $\mathbf{X}$. We aim to capture the correlations between $\mathbf{X}$ and $Y$, $\mathbf{X}$ and $Z$, and the interrelations between $Y$ and $Z$ to simultaneously infer $Y$ and $Z$. To address this problem, we present the *WhoAmI* method, whereby the interrelations between multiple dependent variables can be modeled. As a result, the presented WhoAmI method is able to simultaneously infer users' gender and age. The experiments demonstrate that the proposed method can achieve an accuracy of 80% for predicting users' gender and 73% for predicting users' age according to daily mobile communication patterns, significantly outperforming (by up to 10% in terms of F1-Measure shown in Figure 4.1) several alternative methods (Cf. §4.5 for details of the comparison methods). To scale up the proposed method to handle large-scale networks, we further develop a distributed learning algorithm, which can reduce the computational time to sub-linear speedup ($9 - 10\times$ with 16 CPU cores) by leveraging parallel computing.

We further demonstrate one application scenario of demographic prediction in telecommunication industry. In real world, there are two kinds of mobile subscriptions of a mobile operator: postpaid [230] and prepaid [231]. Specifically, a *postpaid* mobile user is required to create an account by providing detailed demographic information (e.g., name, age, gender, etc.). However, a recent ITU report indicates that there is still a large portion of *prepaid* users (also commonly referred to as pay-as-you-go) who are required to purchase credit in advance of service use. Statistics

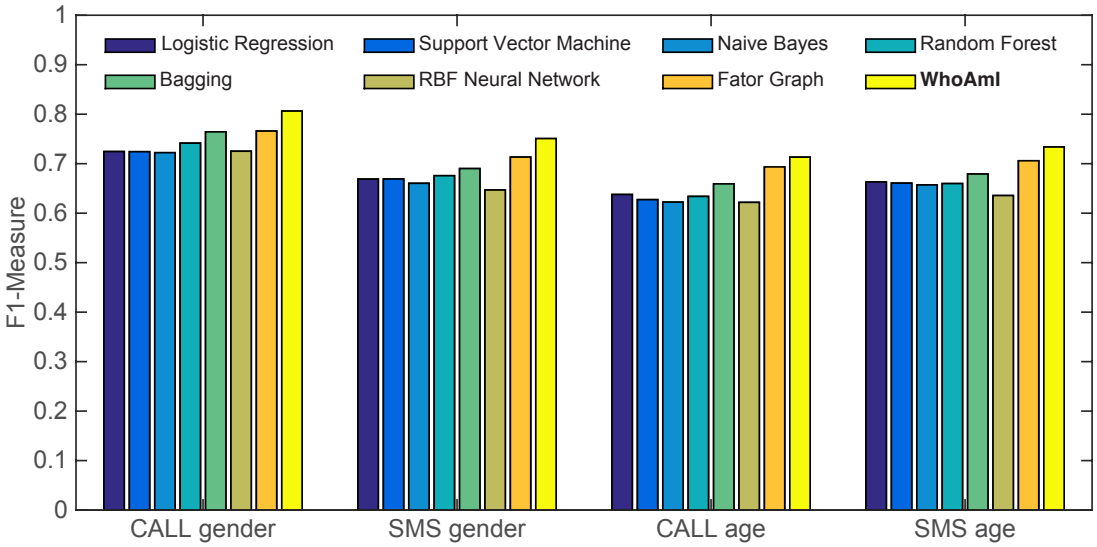



Figure 4.1. Demographic prediction performance. (Cf. §4.5 for details of the comparison methods).

show that 95% of mobile users in India are prepaid, 80% in Latin America, 70% in China, 65% in Europe, and 33% in the United States. Even in the U.S., the switch to prepaid plans was accelerating during the economic recession from 2008. Prepaid services allow the users to be anonymous—no need to provide any user-specific information. In this sense, mobile operators are highly motivated to infer their prepaid users' demographic profiles. We take one case study to demonstrate the effectiveness of our discoveries and methodologies on this real-world application of demographic prediction for prepaid users.

**Coupled Network Demographic Prediction.** In addition to its prepaid users, a mobile operator also does not have the demographic information of users of another operator. For example, in Figure 4.2 a mobile operator $O_1$ (e.g., AT&T) could have the communication logs of two $O_1$ users, and one $O_1$ user and one user of another operator $O_2$ (Verizon) [49]. In real world, $O_1$ does not have the access to the demographic profiles of its competitor $O_2$'s users. However, it is critical for mobile service providers to build the demographic profiles of its competitors' customers. This can

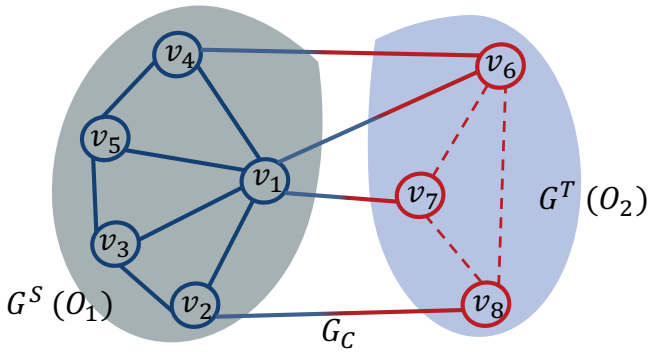



Figure 4.2. An illustrative example of coupled networks across two mobile operators. The source network is mobile operator $O_1$'s network. $O_1$ could also have the demographic information of its own users (postpaid). The objective is to predict the demographic profiles of users in its competitor $O_2$'s network.

help them make better marketing strategies (e.g., identifying potential customers and preventing customer churning). Moreover, by using demographic information, service providers can supply users with more personalized services and focus on enhancing the communication experience.

In light of the real scenario in telecommunication, we formalize the coupled network demographic prediction problem, where we have the structure and user demographic information of one (source) network $G^S$ (e.g., $O_1$) and the interactions between this network and another (target) network $G^T$ (e.g., $O_2$). The goal is to predict the demographic attributes of users in the target network. This problem faces several unique challenges, including the cold start of the target network structure and as a result, the asymmetry of source and target users' graph-based features. To address them, we present a coupled version of the WhoAmI method. Our experiments over six pairs of mobile operators demonstrate the predictability of competitors' user demographics, enabling the potential for business intelligence across mobile operators.

## 4.3 Demographic Prediction Problems

Let $G = (V, E, Y, Z)$ denote the undirected and weighted mobile network, where $V$ is a set of $|V| = N$ users and $E \subseteq V \times V$ is a set of communication edges (CALL or SMS) between users. Each user $v_i \in V$ is associated with demographic information, i.e., gender $y_i \in Y$ and age $z_i \in Z$. We further define an attribute matrix $\mathbf{X}$, where each row $\mathbf{x}_i$ represents an $|\mathbf{x}_i|$ dimensional feature vector for user $v_i$. Given this, we formalize our problem as follows.

**Problem 1** ***Demographic Prediction:*** *Given a partially labeled network* $G = (V^L, V^U, E, Y^L, Z^L)$ *and the attribute matrix* $\boldsymbol{X}$*, where* $V^L$ *is a set of users with labeled demographic information* $Y^L$ *and* $Z^L$*, and* $V^U$ *is a set of unlabeled users, the objective is to learn a function*

$$f : G = (V^L, V^U, E, Y^L, Z^L), \boldsymbol{X} \rightarrow (Y^U, Z^U)$$

*to simultaneously predict users' gender and age, where* $Y^U, Z^U$ *are the demographic information for the unlabeled users* $V^U$*.*

Different from previous work on demographic prediction [17, 95], where users' gender and age are inferred by modeling $P(Y|\mathbf{X})$ and $P(Z|\mathbf{X})$ separately (see Figure 4.3), our problem here is to model $P(Y, Z|G, \mathbf{X})$ for the joint inference of users' gender and age. Specifically, we leverage not only the correlations between $\mathbf{X}$ and $Y/Z$ but also the structural correlations among nodes and interrelations between gender $Y$ and age $Z$. The motivation here comes from the fact that there exist strong network effects and demographic interrelations in human communication behavior, which was demonstrated in previous chapters. For example, a 20-year-old female's behavior is distinct from not only a 20-year-old male's, but also from a 50-year-old female's.

In addition, there are usually multiple mobile operators in telecommunication market—for example, the two mobile operators in Figure 4.2. A mobile operator $O_1$

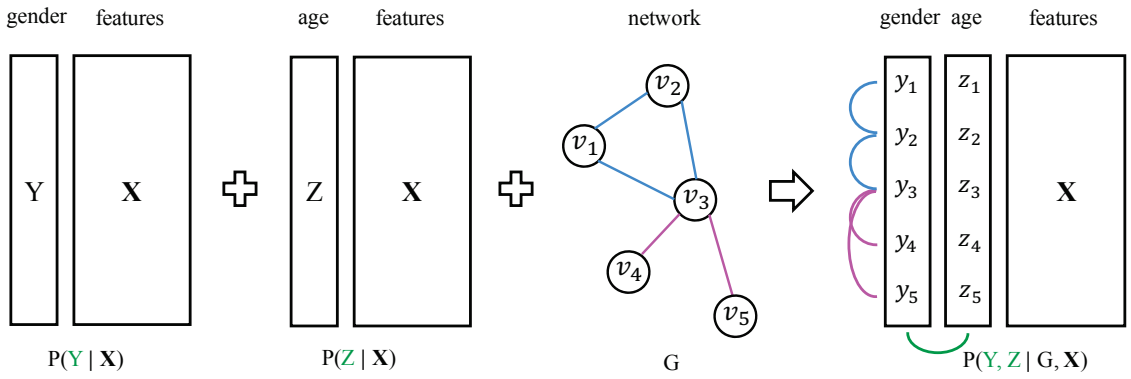



Figure 4.3. An illustration of the proposed demographic prediction problem. In addition to model the correlations between labels ($Y$ or $Z$) and features ($\mathbf{X}$) of each node, we propose to further model the structural correlations among different nodes ($G$) as well as the interrelations between one node's two labels, that is, $Y$ and $Z$.

(e.g., AT&T) could have the communication records of its users and also the communication logs between its users and users of another operator $O_2$ (e.g., Verizon) [49]. It would be very useful for the operator $O_1$ to have the demographic profiles of users of its competitor $O_2$ for business intelligence and precision marketing, such as acquiring new users from and preventing customer churning to competitors.

To solve this problem, we define the concept of coupled networks and formulate the problem of coupled network demographic prediction across multiple operators in mobile communication.

**Definition 1** ***Coupled Networks:*** *Given a source network $G^S = (V^S, E^S)$ and a target network $G^T = (V^T, E^T)$, they compose coupled networks if there exists a cross link $e_{ij}$ with one node $v_i \in V^S$ and the other node $v_j \in V^T$. The cross network $G^C = (V^C, E^C)$ is a bipartite network containing all the cross links in the coupled networks.*

Figure 4.2 shows a typical example of coupled networks with the left network of mobile operator $O_1$ as the source network $G^S$ and the right network of another

mobile operator $O_2$ as the target network $G^T$. The links between these two networks represent the communications between users belonging to these two different mobile operators, which, with their linked nodes in $G^S$ and $G^T$, constitute the cross network $G^C$.

**Problem 2** ***Coupled Network Demographic Prediction:*** *Given the source network $G^S$ with its users' demographic profiles $Y^S, Z^S$ and the cross network $G^C$ in coupled networks $G = (G^S, G^T, G^C)$, the task is to find a predictive function:*

$$f : G^S = (V^S, E^S, Y^S, Z^S), G^C = (V^S, V^T, E^C) \rightarrow (Y^T, Z^T)$$

*where $Y^T$ and $Z^T$ are the set of demographic labels—gender and age—of users $V^T$ in the target network $G^T$.*

The difference between the coupled network demographic prediction and Problem 1 lies in the cold start of network structures between target users in Problem 2. For example, in Figure 4.2, the triangle structures $(v_6, v_7, v_8)$, $(v_1, v_6, v_7)$ can not be observed by the operator $O_1$, making it impossible to leverage the correlations built upon these structures in the prediction task. The real-world and yet challenging setting of the coupled network demographic prediction can be directly applied by a mobile operator to infer the demographic profiles of its competitors' users, facilitating the acquirement of new users from competitor operators.

We treat users' gender as a **binary** random variable, i.e., Female or Male, and users' age as a **four-class** variable by splitting users' age into the four groups mentioned above [17, 95], i.e., Young (18 – 24), Young-Adult (25 – 34), Middle-Age (35 – 49), and Senior ($>$ 49).

## 4.4 The *WhoAmI* Framework

Leveraging the insights gleaned from our network analysis in previous sections, we develop a unified model to capture not only the correlations between users' communication behaviors and demographic profiles but also the interrelations among users' different demographic attributes. In our previous work [44], the proposed *DFG* (Double Label Factor Graph) model is only capable of handling the interrelations between two dependent variables (e.g., gender $Y$ and age $Z$). In this extension, we generalize the WhoAmI method to a Multiple Label Factor Graph Model (MFG). The MFG is general to model the interrelations among multiple (more than two) dependent variables. To illustrate the way that MFG captures the interrelations between multiple ($> 2$) labels, we assume that in addition to one's gender $Y$ and age $Z$, each user is also associated with another demographic attribute $S$ (e.g., income). However, notice that in the mobile data only two demographic attributes—gender and age—are available. Therefore, in Section 4.5 we use the proposed approach to predict these two attributes.

To infer users' demographic attributes in coupled networks, we propose a variant of the Multiple Label Factor Graph—CoupledMFG—that is able to address the unique challenges presented in this task. To handle large-scale networks, we further develop a distributed learning algorithm.

### 4.4.1 Multiple Label Factor Graph

We define an objective function by maximizing the conditional probability of users' gender $Y$, age $Z$, and $S$ given their corresponding attributes $\mathbf{X}$ and the input network structure $G$, i.e., $P_\theta(Y, Z, S|G, \mathbf{X})$. The factor graph [112] provides a way to factorize the "global" probability as a product of "local" factor functions, which

makes the maximization simple, i.e.,

$$P(Y,Z,S|G,\mathbf{X}) = \frac{P(\mathbf{X},G|Y,Z,S)P(Y,Z,S)}{P(\mathbf{X},G)} \propto P(Y,Z,S|G)P(\mathbf{X}|Y,Z,S) \quad (4.1)$$
$$\propto \prod_{v_i \in V} P(\mathbf{x}_i|y_i,z_i,s_i) \prod_{c \in G} P(Y_c,Z_c,S_c)$$

where $P(Y_c, Z_c, S_c)$ denotes the probability of labels given the network structure $c$ and $P(\mathbf{x}_i|y_i, z_i, s_i)$ is the probability of users' attributes $\mathbf{x}_i$ given the labels $y_i$, $z_i$, and $s_i$.

Our proposed model consists of three kinds of factor functions. The first one is an attribute factor $f(y_i, z_i, s_i, \mathbf{x}_i)$ for capturing correlations between users' demographics and communication attributes. The second one is a dyadic factor $g(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)$ for modeling correlations between users' demographics and their direct relationships in ego networks, where $Y_c$ in Eq. 4.1 is represented as $\mathbf{y}_e$ ($y_i$ and $y_j$), $Z_c$ is denoted by $\mathbf{z}_e$ ($z_i$ and $z_j$), and $S_c$ by $\mathbf{s}_e$ ($s_i$ and $s_j$) iff $e_{ij} \in E$. The third one is a triadic factor $h(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)$ for correlating users' demographics and triadic relationships in their ego networks. Similarly, $\mathbf{y}_c$ refers to $y_i, y_j, y_k$, while $\mathbf{z}_c$ refers to $z_i, z_j, z_k$, and $\mathbf{s}_c$ is $s_i, s_j, s_k$ when three users $v_i, v_j, v_k$ form a closed triangle structure $c_{ijk}$, i.e., $e_{ij}, e_{ik}, e_{jk} \in E$.

Therefore, the joint distribution can be further factorized as:

$$P(Y,Z,S|G,\mathbf{X}) = \prod_{v_i \in V} f(y_i,z_i,s_i,\mathbf{x}_i) \times \prod_{e_{ij} \in E} [g(\mathbf{y}_e,\mathbf{z}_e,\mathbf{s}_e)] \times \prod_{c_{ijk} \in G} [h(\mathbf{y}_c,\mathbf{z}_c,\mathbf{s}_c)] \quad (4.2)$$

Figure 4.4 shows an illustration of our proposed model, which consists of two layers of nodes. The bottom layer contains random variables and the upper layer contains the three kinds of factors introduced above. The joint distribution over the whole set of random variables can be factorized as the product of all factors. Specifically, we instantiate the three factors as follows.

**Attribute Factors.** We design the factor $f(y_i, z_i, s_i, \mathbf{x}_i)$ to represent the corre-

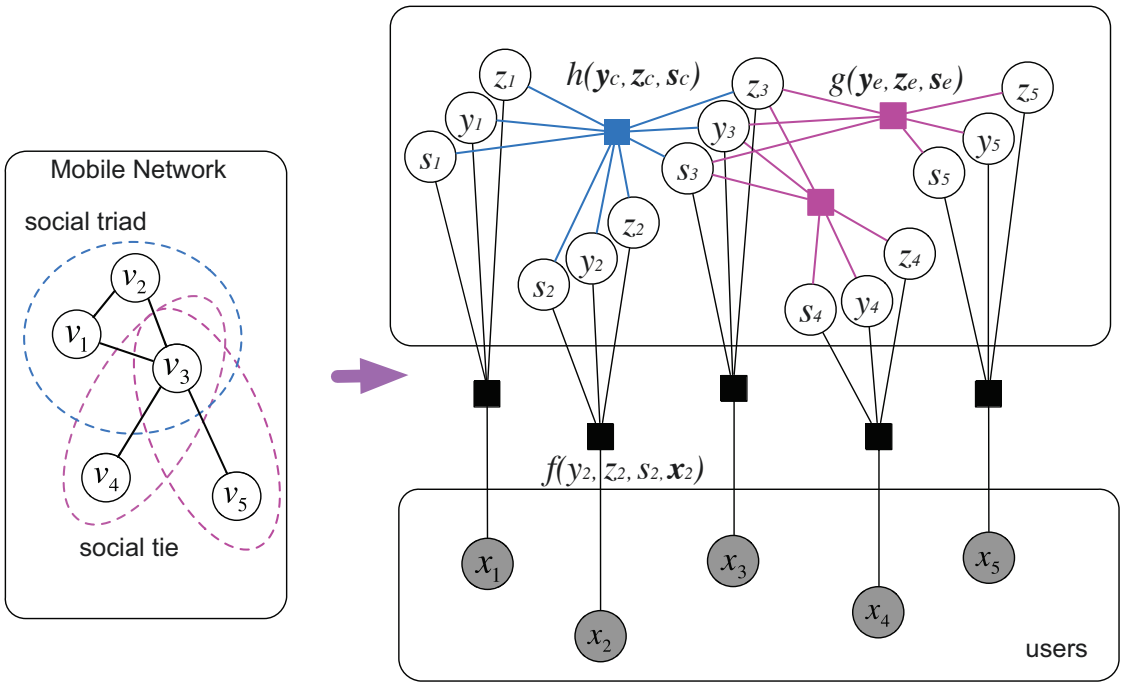



Figure 4.4. An illustration of the proposed WhoAmI model. $y_\cdot$, $z_\cdot$ and $s_\cdot$ indicate the gender, age, and newly added label of the user $v_i$. $x_i$ denotes communication attributes of the user $v_i$ extracted from the mobile network $G$. $f(y_i, z_i, s_i, \mathbf{x}_i)$, $g(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)$, and $h(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)$ respectively represent attribute factor, dyadic factor, and triadic factor in the proposed model.

lation between user $v_i$'s demographics and her/his network characteristics $\mathbf{x}_i$. More specifically, we instantiate the factor by an exponential-linear function:

$$f(y_i, z_i, s_i, \mathbf{x}_i) = \frac{1}{W_v} \exp\{\alpha_{y_i z_i s_i} \cdot \mathbf{x}_i\} \tag{4.3}$$

where $\alpha$ is one parameter of the proposed model, and $W_v$ is a normalization term. For each $(y_i, z_i, s_i)$, $\alpha_{y_i z_i s_i}$ is an $|\mathbf{x}|$-length vector, where the $k$-th dimension indicates how $x_{ik}$ distributes over $(y_i, z_i, s_i)$. For example, let's say $x_{ik}$ represents the degree of user $v_i$. This factor can capture the fact that people with different demographic profiles have different network properties shown in Figure 2.2. Traditional probabilistic graphical models can only model the correlations between features and one single type

of dependent variable, while our proposed model captures how the features jointly distribute over multiple dependent variables.

**Dyadic Factors.** We next define the dyadic factor $g(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)$, where $e_{ij} \in E$, to represent the correlation between user $v_i$ and $v_j$'s demographic information. Specifically, we have

$$g(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e) = \begin{cases} \frac{1}{W_{e_1}} \exp\{\beta_1 \cdot g'_1(y_i, y_j)\} \\ \frac{1}{W_{e_2}} \exp\{\beta_2 \cdot g'_2(y_i, z_i)\} \\ \dots \\ \frac{1}{W_{e_{14}}} \exp\{\beta_{14} \cdot g'_{14}(z_j, s_i)\} \\ \frac{1}{W_{e_{15}}} \exp\{\beta_{15} \cdot g'_{15}(s_i, s_j)\} \end{cases} \tag{4.4}$$

where $\beta_p$ is the model parameter for this type of factor, $g'_p(\cdot)$ is defined as a vector of indicator functions, and $W_{e_p}$ is the normalization term. We can enumerate in total $C_6^2 = 15$ different combinations of each pair of demographic variables from $(y_i, y_j, z_i, z_j, s_i, s_j)$. The intuition behind this is that $v_i$'s friends' demographics distribute differently by varying either $v_i$'s own age or gender or income, as Figure 2.3 suggests.

**Triadic Factors.** We finally define the triadic factor $h(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)$ to represent the correlation among the demographics of social triads, where $c = \{v_i, v_j, v_k | e_{ij}, e_{jk}, e_{ik} \in$

$E\}$ indicates the closed triangle structure in $G$. More specifically, we have

$$h(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c) = \begin{cases} \frac{1}{W_{c_1}} \exp\{\gamma_1 \cdot h'_1(y_i, y_j, y_k)\} \\ \frac{1}{W_{c_2}} \exp\{\gamma_2 \cdot h'_2(y_i, y_j, z_i)\} \\ \dots \\ \frac{1}{W_{c_{83}}} \exp\{\gamma_{83} \cdot h'_{83}(s_i, s_j, z_k)\} \\ \frac{1}{W_{c_{84}}} \exp\{\gamma_{84} \cdot h'_{84}(s_i, s_j, s_k)\} \end{cases} \tag{4.5}$$

where $h'_q(\cdot)$ is the vector of indicator functions and $W_{c_q}$ is the normalization term similar with $W_{e_p}$. There are $C_9^3$ different kinds of three-variable enumerations from $(y_i, y_j, y_k, z_i, z_j, z_k, s_i, s_j, s_k)$. We use these triadic factors to model the distributions of user demographics within a closed social triangle (see details in Figure 2.5).

Finally, incorporating Eqs. 4.3, 4.4, 4.5 into Eq. 4.2, we define the objective function as the log-likelihood of the proposed model as:

$$\mathcal{O}(\alpha, \beta, \gamma) = \sum_{v_i \in V} \alpha_{y_i z_i s_i} \mathbf{x}_i + \sum_{e_{ij} \in E} \sum_{p=1}^{15} \beta_p g'_p(\cdot) + \sum_{c_{ijk} \in G} \sum_{q=1}^{84} \gamma_q h'_q(\cdot) - \log W \tag{4.6}$$

where $W = W_v W_e W_c$ is the global normalization term, $W_e = \prod_{e_p=1}^{15} W_{e_p}$, and $W_c = \prod_{c_q=1}^{84} W_{c_q}$.

The technical novelty of the proposed model is that it considers different types of labels in a unified framework, which differentiates our model from traditional classification models. By considering three types of labels in this special case, the main advantage is that our model can characterize the interrelations between different demographic labels and the structural correlations between different users as well as correlations between labels and features.

### 4.4.2 Feature Definition

Given a network with labeled and unlabeled users, the goal is to infer unlabeled users' demographic information, which is in accordance with the real-world application scenarios. There are two types of features designed in our experiments, namely *nonstructural attribute features* and *structural features*. Specifically, given an ego network with one central user $v$ and her/his direct friends, we extract three kinds of attribute features for this central user $v$ as follows:

*Individual attributes* are extracted based on the network topological properties discussed in ego networks. It includes the degree, neighbor connectivity, clustering coefficient, embeddedness, and weighted degree (#calls or #messages) of each node.

*Friend attributes* are used to model the demographic distribution of $v$'s direct friends in her/his ego network, including the number of connections to female, male, young, young-adult, middle-age, and senior friends. In the prediction setting, not all friends of the central user $v$ are labeled with gender or age information, so we extract the friend attributes only based on her/his labeled friends.

*Circle attributes* refer to the triadic demographic distribution of $v$'s ego network. Because we aim to infer the central user $v$'s demographics, we count the numbers of different gender triads, i.e., '$FF$-$v$', '$FM$-$v$', '$MM$-$v$', and different age-group triads. Let A/B/C/D denote the young/young-adult/middle-age/senior age-groups, respectively. There are in total ten kinds of triads based on age-groups: '$AA$-$v$', '$AB$-$v$', '$AC$-$v$', '$AD$-$v$', '$BB$-$v$' ,'$BC$-$v$' ,'$BD$-$v$', '$CC$-$v$', '$CD$-$v$', '$DD$-$v$'.

Table 4.1 lists 24 nonstructural attribute features used in our models. Notice that friend and circle attributes can only be extracted from $v$'s labeled friends. These three types of attribute features—individual, friend, and circle attributes — are captured by the attribute factor in our MFG model (Cf. Eq. 4.3).

In addition, the structural features, captured by the dyadic factor (Cf. Eq. 4.4) and triadic factor (Cf. Eq. 4.5), are designed to model the demographic distributions

over edges and triangles with both labeled and unlabeled users, which forms one of the advantages of the proposed factor graph-based model. Together with nonstructural friend attributes, structural features covered by dyadic factors form *friend features*. Similarly, *circle features* are composed of nonstructural circle attributes and triadic structural features.

### 4.4.3 Learning and Inference

The goal of learning the WhoAmI method is to find a configuration for the free parameters $\theta = \{\alpha, \beta, \gamma\}$ that maximize the log-likelihood of the objective function $\mathcal{O}(\theta)$ in Eq. 4.6 given by the training set, i.e., $\theta^\star = \arg\max \mathcal{O}(\theta)$.

**Learning.** We first introduce how we learn the model in a single-processor configuration, and then explain how to extend the learning algorithm to a distributed one for handling large-scale networks.

To solve the optimization problem, we adopt a gradient decent method (or a Newton-Raphson method). Specifically, we derive the objective function with respect to each parameter with regard to our objective function in Eq. 4.6.

$$\begin{aligned}
\frac{\partial \mathcal{O}(\theta)}{\partial \alpha} &= \boldsymbol{E}[\sum_{v_i \in V} \mathbf{x}_i] - \boldsymbol{E}_{P_\alpha(Y,Z,S|\mathbf{X})}[\sum_{v_i \in V} \mathbf{x}_i] \\
\frac{\partial \mathcal{O}(\theta)}{\partial \beta} &= \boldsymbol{E}[\sum_{e_{ij} \in E} g'(\cdot)] - \boldsymbol{E}_{P_\beta(Y,Z,S|\mathbf{X},G)}[\sum_{e_{ij} \in E} g'(\cdot)] \\
\frac{\partial \mathcal{O}(\theta)}{\partial \gamma} &= \boldsymbol{E}[\sum_{c_{ijk} \in G} h'(\cdot)] - \boldsymbol{E}_{P_\gamma(Y,Z,S|\mathbf{X},G)}[\sum_{c_{ijk} \in G} h'(\cdot)]
\end{aligned} \tag{4.7}$$

where in the first Equation of Eq. 4.7, $\boldsymbol{E}[\sum_{v_i \in V} \mathbf{x}_i]$ is the expectation of the summation of the attribute factor functions given the data distribution over $Y$, $Z$, $S$, and $\mathbf{X}$ in the training set, and $\boldsymbol{E}_{P_\alpha(Y,Z,S|X)}[\sum_{v_i \in V} \mathbf{x}_i]$ is the expectation of the summation of the attribute factor functions given by the estimated model. The other expectation terms have similar meanings in the other two equations. As the net-

work structure in the real-world may contain cycles, it is intractable to estimate the marginal probability in the second terms of Eq. 4.7. In this work, we adopt Loopy Belief Propagation (LBP) [158] to calculate the marginal probability of $P(Y, Z, S)$ and compute the expectation terms.

The learning process then can be described as an iterative algorithm. Each iteration contains two steps: First, we call LBP to calculate marginal distributions of unknown variables $P_\alpha(Y, Z, S|X)$. Second, we update $\alpha$, $\beta$, and $\gamma$ with the learning rate $\eta$ by Eq. 4.8. The learning algorithm terminates when it reaches convergence.

$$\theta_{new} = \theta_{old} + \eta \cdot \frac{\partial \mathcal{O}(\theta)}{\partial \theta} \tag{4.8}$$

**Prediction.** With the estimated parameter $\theta$, we can now assign the value of unknown labels $Y, Z, S$ by looking for a label configuration that will maximize the objective function, i.e. $(Y^*, Z^*, S^*) = \arg\max\ \mathcal{O}(Y, Z, S|G, \mathbf{X}, \theta)$. In this work, we use the max-sum algorithm [112] to solve the above problem.

**Complexity.** The complexity of the learning algorithm at each iteration is $O(|V| \cdot Q + |E| \cdot Q^2 + |C| \cdot Q^3)$, where $|V|, |E|, |C|$ are the numbers of users, edges, and triads in the graph, respectively, and $Q$ is the number of classes of multiple labels. Specifically, $Q = |Y| \times |Z| \times |S|$ in the presented model, where $|Y| = 2$ is the number of gender labels—male and female, $|Z| = 4$ is the number of age labels—young, young-adult, middle-age, and senior, and $|S|$ is the number of income labels. Therefore, when learning over only gender and age in our prediction experiments, $Q$ is equal to $|Y| \times |Z|$, that is 8.

### 4.4.4 Distributed Learning

We further leverage a distributed framework [205, 208] to scale up our model to handle these large-scale mobile networks. Our distributed learning algorithm utilizes

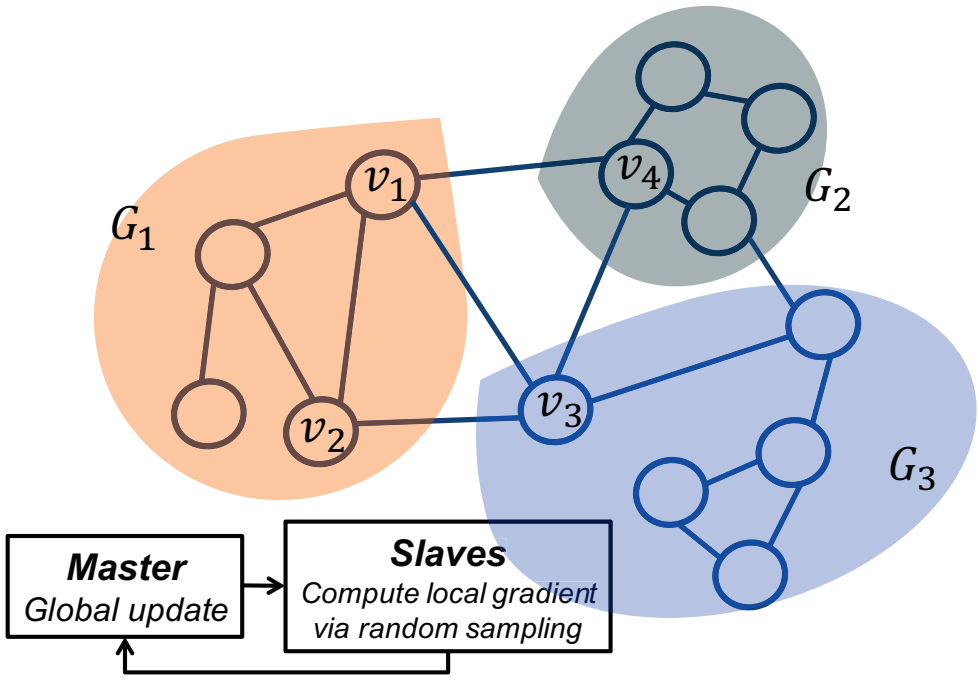



Figure 4.5. An illustration of the master-slave learning scheme.

a Message Passing Interface (MPI) framework, by which we can split the network into small parts and learn the parameters on different processors. As most computing time is consumed in the first step of our learning algorithm introduced above, we speed up this learning process by distributing multiple 'slave' computing processors for this step. The second step is calculated in the 'master' processor by collecting the results from all 'slave' processors on the first step. An illustrative flow of the two steps can be found in Figure 4.5.

Specifically, the master-slave based distributed learning framework [205, 208] can be described in two phases. At the first phase, the large-scale network $G$ is partitioned into $K$ sub-networks $G_1, \cdots, G_k, \cdots, G_K$ of balanced size, and the $K$ sub-networks are distributed to $K$ 'slave' processors. At the second phase, we iteratively learn the parameters in two steps. At each iteration $i$, first, each processor can compute the

local belief on its sub-network $G_k$ according to Eq. 4.9.

$$M_t^{k,i}(\chi_t) \propto f^k(\chi_t, \cdot) \prod_{u \in \Gamma(t)} m_{u \to t}^{k,i}(\chi_t) \tag{4.9}$$

where $\chi_t$ denotes the nodes in the local factor graph, $\Gamma(t)$ denotes $\chi_t$'s neighbors, and $m_{u \to t}^{k,i}$ denotes the belief (message) propagated from node $\chi_u$ to node $\chi_t$, which is defined as the following equation.

$$m_{u \to t}^{k,i}(\chi_t) \propto \sum_{\chi_u} f^k(\chi_u, \cdot) g^k(\chi_u, \chi_t) h^k(\chi_u, \chi_t, \cdot) \prod_{s \in \Gamma(u) \setminus t} m_{s \to u}^{k,i-1}(\chi_u) \tag{4.10}$$

wherein the message will be normalized. Second, the 'master' processor collects all local results obtained from different subgraphs and computes the marginal probability $P(\chi_t|\cdot)$ according to Eq. 4.11, and updates parameters according to Eqs. 4.7 and 4.8.

$$P^i(\chi_t|\cdot) = \sigma \sum_{k=1}^{K} M_t^{k,i}(\chi_t) \tag{4.11}$$

where $\sigma$ is the normalization constant. This phase is repeated until convergence.

There are three notes for our model implementation. In order to achieve the balance among different slaves, we partition the nationwide mobile network into $K$ subgraphs of roughly equal size. The second one is that we first extract all features for each user from the original full network. We then split it into subgraphs that are handled by each 'slave' processor.

The third point worth noting is that a structural factor has to be eliminated in the distributed learning framework if it is defined over several nodes that belong to different subgraphs—for example, the triangle structures $(v_1, v_2, v_3)$ and $(v_1, v_3, v_4)$ in Figure 4.5. To address this issue, we propose to use virtual nodes [205, 208] to construct the broken structural factors. For example, to complete the triad factor over the triangle $(v_1, v_2, v_3)$ that would be ignored in $G_1$ in Figure 4.5, we design a

virtual node $v_3'$ in $G_1$. In doing so, the factor graph over $G_1$ will capture the structural correlations of the three users' demographic information. As the completion of the triangle $(v_1, v_2, v_3)$ in $G_1$, it will not be constructed in the other subgraph, that is, $G_3$. With that said, if three nodes of a triangle are distributed into three subgraphs, such as $(v_1, v_3, v_4)$, one of the three involved subgraphs will be randomly selected to complete the triangle and leave the other two ignored.

### 4.4.5 Coupled Network Learning

Finally, we design a variant of the WhoAmI method to address the challenges in coupled network demographic prediction. As illustrated in Section 4.3, the problem faces two unique challenges. First, the missing of the target network structure makes it impossible to define triadic factors $h(\cdot)$ over three target users, such as the triangle structure $(v_6, v_7, v_8)$ in Figure 4.2. Second, users' individual features across different mobile operators are asymmetric, due to the sparsity of the target network. For example, the connections between user $v_1$ and users from both the same operator $O_1$ $(v_2, v_3, v_4, v_5)$ and the other operator $O_2$ $(v_6, v_7)$ are observed for counting $v_1$'s degree centrality, while for user $v_6$ in $O_2$, the associations with $O_1$'s users $(v_1, v_4)$ can be observed, and those with target users $(v_7, v_8)$ are not observable. In this context, the individual features of source and target users follow different distributions, making it infeasible for a supervised learning framework.

In light of these issues and our previous work on coupled link prediction [49], we propose the coupled version of the WhoAmI method—CoupledMFG. By taking the coupled mobile networks as the input of a factor graph model, we have the following

joint distribution:

$$P(Y, Z, S|G^S, G^C, \mathbf{X}) = \prod_{v_i \in V^S} f^S(y_i, z_i, s_i, \mathbf{x}_i) \times \prod_{v_i \in V^T} f^T(y_i, z_i, s_i, \mathbf{x}_i) \times \prod_{e_{ij} \in E^S} [g^S(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)] \times \prod_{e_{ij} \in E^C} [g^C(\mathbf{y}_e, \mathbf{z}_e, \mathbf{s}_e)] \times \prod_{c_{ijk} \in G^S} [h^S(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)] \times \prod_{c_{ijk} \in G^C} [h^C(\mathbf{y}_c, \mathbf{z}_c, \mathbf{s}_c)] \tag{4.12}$$

This joint distribution factorizes all factors over the available structures in coupled networks. The first two terms model the attribute factors for users in source and target networks, respectively. Recall that one of the challenges is the asymmetry of users' individual attributes across these two networks, making it desired to separately model these two groups of attribute factors $f^S(\cdot)$ and $f^T(\cdot)$. The remaining four terms capture the structural correlations in coupled networks. Specifically, the third and fourth terms model the dyadic correlations, and the fifth and sixth terms model the triadic correlations in the source and cross networks, respectively. Further, all the latent variables in $g^S(\cdot)$ and $h^S(\cdot)$ are labeled, while only partial of latent variables in $g^C(\cdot)$ and $h^C(\cdot)$ are known to the model. Take the triadic factor $h^C(\cdot)$ over the triangle $(v_1, v_4, v_6)$ in Figure 4.2 as an example, user $v_6$'s demographic attributes are not available—in fact, they are the objective of the prediction model—and the demographics of users $v_1$ and $v_4$ are labeled for the learning algorithm.

One necessary question arises: Do the demographic correlations over edges $g(\cdot)$ and triangles $h(\cdot)$ follow the same distribution in source and cross networks? Our examination shows that there exists no significant distinction on the demographic distributions between source and cross networks. With that said, the semi-supervised nature of the proposed WhoAmI method enables the joint modeling of structural factors ($g(\cdot)$ and $h(\cdot)$) across source and target networks. To do so, we model the structural factors into the same parameter space. Specifically, we have the following

**ALGORITHM 1:** Distributed CoupledMFG Learning Algorithm.

**Input:** The source network $G^S$, the cross network $G^C$, the node set $V^T$ of the target network $G^T$, and the learning rate $\eta$

**Output:** Parameters $\theta = (\alpha^S, \alpha^T, \beta, \gamma)$

Master initializes $\theta \leftarrow 0$;
Master constructs the coupled factor graph according to Eq. 4.12 with $G^S, G^C, V^T$;
Master partitions the input mobile network into $K$ subgraphs of relatively equal size;
Master completes the broken structural factors with virtual nodes;
Master forwards all subgraphs to slaves [Communication];
**repeat**
  Master broadcasts $\theta$ to Slaves [Communication];
  **for** $k = 1 \rightarrow K$ **do**
    Slave $k$ computes local belief according to Eqs. 4.9 and 4.10;
    Slave $k$ sends the local belief to Master [Communication];
  **end**
  Master calculates the marginal distribution for each variable according to Eq. 4.11;
  Master calculates the gradient for each parameter according to Eq. 4.7;
  Master updates the parameters according to Eq. 4.8;
**until** *Convergence*;

log-likelihood objective function for the CoupledMFG model.

$$\mathcal{O}(\alpha, \beta, \gamma) = \sum_{v_i \in V^S} \alpha^S_{y_i z_i s_i} \mathbf{x}^S_i + \sum_{v_i \in V^T} \alpha^T_{y_i z_i s_i} \mathbf{x}^T_i + \sum_{p=1}^{15} \beta_p \sum_{e_{ij} \in E^S \cup E^C} g'_p(\cdot) + \sum_{q=1}^{84} \sum_{c_{ijk} \in G^S \cup G^C} \gamma_q h'_q(\cdot) - \log W \tag{4.13}$$

where the two different parameters $\alpha^S$ and $\alpha^T$ are designed to separately model the attribute factors in source and target networks, and on the other hand, both the parameters $\beta$ and $\gamma$ are used to simultaneously model the dyadic and triadic factors across source and cross networks. In doing so, the CoupledMFG model is enabled to handle the two challenges in coupled network demographic prediction—the sparseness of the target network and as a result, the asymmetry of individual features in source

and target networks.

The distributed learning algorithm for CoupledMFG is presented in Algorithm 1. In the algorithm, we also mark the communications between Master and Slaves. The learning algorithm will assign the target users (unlabeled) with demographic labels that maximize the marginal probabilities.

## 4.5 Experiments

We present the effectiveness and efficiency of our proposed WhoAmI method on demographic prediction by various experiments. The code used in the experiment is publicly available at `http://arnetminer.org/demographic`.

### 4.5.1 Experiment Setup

**Data and Evaluation.** We use two large-scale mobile networks, CALL and SMS, to infer users' gender and age. To infer user demographics effectively for mobile operators, we only consider active users who have at least five contacts in two months. After filtering out non-active users, there are 1.09 million and 304,000 active users in the CALL and SMS networks, respectively. We repeat the prediction experiments ten times, and report the average performance in terms of weighted Precision, Recall, and F1-Measure. We consider weighted evaluation metrics because every class in female / male or young / young-adult / middle-age / senior is as important as each other.

All code is implemented in C++, and prediction experiments are performed in a server with four 16-core 2.4 GHz AMD Opteron processors with 256GB RAM. We use the speedup metric with different numbers of computing cores (1-16) to evaluate the scalability of our distributed learning algorithm.

**Comparison Methods.** We compare our proposed WhoAmI method that can capture the interrelation between two types of labels (gender and age) with different

classification algorithms, including Logistic Regression (**LRC**), Support Vector Machine (**SVM**), Naive Bayes (**NB**), Random Forest (**RF**), Bagging (**Bag**), Gaussian Radial Basis Function Neural Network (**RBF**), and Factor Graph Model (**FGM**). For LRC, NB, RF, Bag, RBF, we employ Weka [84] and use the default setting and parameters. For SVM, we use liblinear [63]. For FGM, the model proposed in [130] is used. Note that our proposed WhoAmI method is equal to FGM if we do not consider the interrelations between gender and age. In addition, other types of models have been used for capturing interaction effects from data, such as hierarchical multi-level models [71, 172]. However, rather than detecting and modeling the nested structures, the goal of this work is to demonstrate the effects of dyadic and triadic correlations between users' demographic attributes. Therefore, those models are not considered in the experiments.

For all comparison methods, we use the same unstructured features (individual, friend, and circle attributes) introduced in Feature Definition of Section 4.4.2. For the graphical models, FGM and WhoAmI, the structural features (dyadic and triadic factors) are further used to model user demographics on network structure. The major difference between our WhoAmI method and the FGM model is that WhoAmI can capture not only the structural correlations between different users, but also the interrelations between two dependent variables of each user, i.e., gender and age.

### 4.5.2 Experiment Results

We report the demographic prediction performance for different methods in the CALL and SMS networks. In prediction experiments, we use 50% of the labeled data in each network as training set and the remaining 50% for testing.

**Predictive Performance.** Tables 4.2 and 4.3 show the prediction results of different algorithms on the four prediction cases, i.e., gender and age predictions in the CALL and SMS networks, respectively. Clearly our WhoAmI method yields better

performance than the other alternative methods in all four cases. The Bag method achieves the best prediction results among all non-graphical methods. The FGM model outperforms a series of non-graphical algorithms by modeling the correlations among structured nodes via dyadic and triadic factors. The WhoAmI method outperforms FGM by further leveraging the interrelations between users' gender and age. In terms of weighted Precision, Recall, and F1-Measure, WhoAmI achieves up to 10% improvements compared with the baselines for the prediction of users' gender and age. As for Accuracy, the WhoAmI method can infer 80% of the users' gender in the CALL network and 73% of the users' age in the SMS network correctly. Finally, we observe that the CALL network can reveal more users' gender information than the SMS network, as the overall performance of gender prediction in CALL is about 5% higher than that in SMS. However, predicting age from text messaging behavior is relatively easier than predicting it from phone call communications. The reason can be reasoned from the discoveries in Section 2.4, where we find that the difference on the usage of text messages between the young and senior people is more strong than that in phone call usage, resulting the better performance in age prediction in SMS than CALL, while the gender homophily in phone calls is more obvious than in messages, leading to the advantage when predicting gender from the CALL network.

**Effects of Demographic Interrelations.** We evaluate the effects of demographic interrelation on the predictions. Without modeling the interrelation between gender and age, our proposed WhoAmI method degenerates to a basic factor graph model. From Tables 4.2 and 4.3 , we clearly observe the 2% to 4% improvements achieved by WhoAmI to FGM on weighted F1-Measure. We further analyze feature contributions for demographic prediction. Recall that in Feature Definition of Section 4.4.2, besides the individual features, we introduced the friend features (nonstructural friend attributes and dyadic factors) and circle features (nonstructural circle attributes and triadic factors). By removing either friend or circle features, we evaluate the de-

crease in predictive performance in terms of weighted F1-Measure, plotted in Figure 4.6. WhoAmI-df, WhoAmI-dc, and WhoAmI-dfc stand for the removing of friend features, circle features, and both of them, conditioned on WhoAmI-d without modeling gender and age interrelations. Clearly, we can see that for inferring gender, the performance when removing circle features drops more than when removing friend features, which indicates a stronger contribution of circle features to gender prediction than friend features. However, for inferring users' age, friend features are more telling than circle features. The feature contribution analysis further validates our observations of demographic-based social strategies, and demonstrates that the proposed model works well by capturing the observed phenomena.

TABLE 4.1

DEFINITION OF NONSTRUCTURAL ATTRIBUTE FEATURES

| Attribute Type | Name | Description |
|---|---|---|
| Individual | degree | number of contacts |
| | neighbor connectivity | average degree of neighbors |
| | clustering coefficient | local clustering coefficient |
| | embeddedness | common neighbor connectivity |
| | weighted degree | #communications |
| Friend | #female-friends | #female contacts |
| | #male-friends | #male contacts |
| | #young-friends | #young contacts |
| | #young-adult-friends | #young-adult contacts |
| | #middle-age-friends | #middle-age contacts |
| | #senior-friends | #senior contacts |
| Circle<br>A: young<br>B: young-adult<br>C: middle-age<br>D: senior | #v-FF-triangles | #$FF$-$v$ triangles in $v$'s ego network |
| | #v-FM-triangles | #$FM$-$v$ triangles in $v$'s ego network |
| | #v-MM-triangles | #$MM$-$v$ triangles in $v$'s ego network |
| | #v-AA-triangles | #$AA$-$v$ triangles in $v$'s ego network |
| | #v-AB-triangles | #$AB$-$v$ triangles in $v$'s ego network |
| | #v-AC-triangles | #$AC$-$v$ triangles in $v$'s ego network |
| | #v-AD-triangles | #$AD$-$v$ triangles in $v$'s ego network |
| | #v-BB-triangles | #$BB$-$v$ triangles in $v$'s ego network |
| | #v-BC-triangles | #$BC$-$v$ triangles in $v$'s ego network |
| | #v-BD-triangles | #$BD$-$v$ triangles in $v$'s ego network |
| | #v-CC-triangles | #$CC$-$v$ triangles in $v$'s ego network |
| | #v-CD-triangles | #$CD$-$v$ triangles in $v$'s ego network |
| | #v-DD-triangles | #$DD$-$v$ triangles in $v$'s ego network |

TABLE 4.2

CALL DEMOGRAPHIC PREDICTION PERFORMANCE

| Network | Method | Gender | | | Age | | |
|---|---|---|---|---|---|---|---|
| | | wPrecision | wRecall | wF1-Measure | wPrecision | wRecall | wF1-Measure |
| CALL | LRC | 0.7327 | 0.7289 | 0.7245 | 0.6350 | 0.6466 | 0.6337 |
| | SVM | 0.7327 | 0.7287 | 0.7242 | 0.6369 | 0.6463 | 0.6273 |
| | NB | 0.7222 | 0.7227 | 0.7222 | 0.6246 | 0.6224 | 0.6223 |
| | RF | 0.7437 | 0.7310 | 0.7415 | 0.6382 | 0.6482 | 0.6388 |
| | Bag | 0.7644 | 0.7648 | 0.7643 | 0.6607 | 0.6688 | 0.6592 |
| | RBF | 0.7283 | 0.7275 | 0.7252 | 0.6194 | 0.6272 | 0.6218 |
| | FGM | 0.7658 | 0.7662 | 0.7659 | 0.6998 | 0.6989 | 0.6935 |
| | WhoAmI | 0.8088 | 0.8076 | 0.8063 | 0.7266 | 0.7140 | 0.7132 |

TABLE 4.3

SMS DEMOGRAPHIC PREDICTION PERFORMANCE

| Network | Method | Gender | | | Age | | |
|---|---|---|---|---|---|---|---|
| | | wPrecision | wRecall | wF1-Measure | wPrecision | wRecall | wF1-Measure |
| SMS | LRC | 0.6766 | 0.6758 | 0.6689 | 0.6702 | 0.6890 | 0.6630 |
| | SVM | 0.6749 | 0.6750 | 0.6690 | 0.6654 | 0.6884 | 0.6607 |
| | NB | 0.6231 | 0.6655 | 0.6603 | 0.6563 | 0.6588 | 0.6570 |
| | RF | 0.6399 | 0.6749 | 0.6757 | 0.6623 | 0.6775 | 0.6598 |
| | Bag | 0.6905 | 0.6918 | 0.6901 | 0.6907 | 0.6987 | 0.6791 |
| | RBF | 0.6712 | 0.6592 | 0.6468 | 0.6295 | 0.6640 | 0.6356 |
| | FGM | 0.7132 | 0.7138 | 0.7133 | 0.7154 | 0.7154 | 0.7059 |
| | WhoAmI | 0.7589 | 0.7549 | 0.7507 | 0.7409 | 0.7303 | 0.7337 |

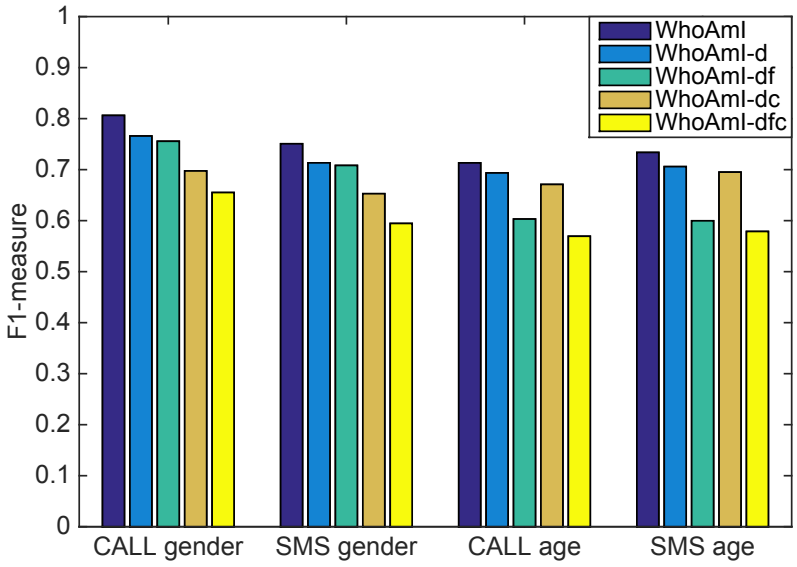
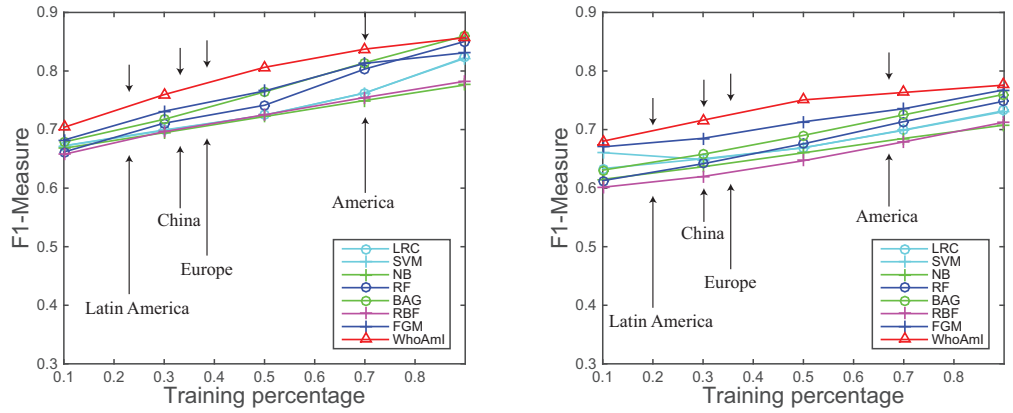



Figure 4.6. Feature Contribution Analysis. WhoAmI is the proposed model. WhoAmI-d is the basic version of WhoAmI without modeling the correlation between gender and age. WhoAmI-df stands for further ignoring friend features. WhoAmI-dc stands for further ignoring circle features. WhoAmI-dcf stands for ignoring both friend and circle features.

**Scalability.** We verify the distributed learning algorithm by partitioning the original large-scale networks into multiple sub-networks based on users' administrative areas. Users' areas are determined by their postal codes during subscription registration. Each sub-network in one area is used as the input for a given core. By utilizing MPI, our distributed algorithm can achieve $9 - 10\times$ speedup with 16 cores with less than 2% drop in performance. Basically, our learning algorithm can converge in 100 iterations, and each iteration costs about 2 (SMS) or 5 minutes (CALL) for one single processor. By leveraging a distributed learning algorithm, our WhoAmI model is efficient even for large-scale networks with millions of nodes.

**Application—Predicting Prepaid Users.** As introduced before, mobile operators may not have the demographic information of prepaid users, and the percentages of prepaid users in mobile operators of different countries are different, such as 95% in




(a) CALL Gender Prediction

(b) SMS Gender Prediction

(c) CALL Age Prediction

(d) SMS Age Prediction

Figure 4.7. Application. Performance of demographic prediction with different percentages of postpaid users.

India, 80% in Latin America, 70% in China, 65% in Europe, and 33% in America. We use different ratios of users as training data and the remaining as testing data. In this way, we can simulate the effects of different percentages of prepaid users on predictive performance. Figure 4.7 shows the prediction results when varying the percentage of labeled users in the training set. Clearly, we can see rising trends as the training set increases in Figure 4.7(a) and 4.7(b). This indicates the positive effects of training data size on predicting the gender of mobile users. Specifically, we can see that in this simulation, the performance for predicting the gender of prepaid users can reach

$\sim$70% in India (5% users as training) in terms of weighted F1-Measure, $\sim$75% in China (30% users as training), and $\sim$83% in America (67% users as training). The smooth lines in Figure 4.7(c) and 4.7(d) reveal the limited contributions of training data size on predicting age. We can see that in all cases, obvious improvements can be obtained by our proposed WhoAmI method with different sizes of training data.

### 4.5.3 Coupled Network Demographic Prediction

We further study how the coupled variant of the WhoAmI method can be used by a mobile operator to infer the demographic profiles of its competitors' users. As the example illustrated in Figure 4.2, a mobile operator $O_1$ could have the communication records of its users and also the communication logs between its users and users of another operator $O_2$ [49]. It would be very useful for the operator $O_1$ to have the demographic profiles of users of the competitor $O_2$ for business intelligence.

In this mobile dataset, there are three major mobile operators. We denote each of the three operators as $O_0$, $O_1$, and $O_2$, respectively. Tables 4.4 and 4.5 list the numbers of active users in the CALL and SMS networks of each operator, and the numbers of edges within ($O_i \rightarrow O_i$) and across different operators ($O_i \rightarrow O_j$). We train the coupled WhoAmI model by taking one operator's network as the source network and another one's as the target network. In total, we construct six pairs of prediction cases in the CALL and SMS networks, respectively, that is, $O_0$ to $O_1$, $O_0$ to $O_2$, $O_1$ to $O_0$, $O_1$ to $O_2$, $O_2$ to $O_0$, and $O_2$ to $O_1$.

Table 4.6 shows the strong predictability of users' demographic attributes across each pair of mobile operators. In general we can see that the predictive performance is very promising compared to the results in Tables 4.2 and 4.3 . Specifically, the results demonstrate that the coupled WhoAmI method offers a 67% $\sim$ 80% predictability for inferring competitor users' gender and a greater than 65% potential for the inference of their age. In other words, a mobile operator would know the demographic profiles

of as many as more than half of its competitors' users, enabling the real-world application of business intelligence in telecommunication, such as acquiring new users from competitors through precision marketing.

We also notice that the prediction cases with a larger mobile operator (more users) as the training data and a smaller operator as the targeting data perform better than those with them exchanged, i.e., the cases $O_0$ to $O_1$, $O_0$ to $O_2$, and $O_1$ to $O_2$ outperform the cases $O_1$ to $O_0$, $O_2$ to $O_0$, and $O_2$ to $O_1$, where the size $|O_0| > |O_1| > |O_2|$. Recall that the coupled prediction task is set in real-world scenarios (Cf. Figure 4.2), that is, the source operator can only observe partial information about the target network, making it infeasible to compute the user distribution distances between its users and target operator users. However, to reason about the outperformance when predicting from $O_{large}$ to $O_{small}$, we report the average number of connections of users from each operator in Tables 4.4 and 4.5. In a composite network of two operators, such as $O_0$ (large) and $O_1$ (small), $O_1$ users on average have more $O_0$ connections than $O_1$ connections (1.82 vs. 1.45 in CALL and 1.41 vs. 1.06 in SMS). In other words, users in a small operator associate more with users of a large operator than users of the same operator. Not surprisingly, users in the large operator $O_0$ have higher rates of same-operator contacts than of $O_1$ connections (2.12 vs. 0.88 in CALL and 1.59 vs. 0.76 in SMS). Consequently, the large operator $O_{large}$ is able to collect rich structural information about target users from its competitors $O_{small}$ who have smaller user base, due to those targets communicate more intensively with $O_{large}$ users than themselves—$O_{small}$. This enables its advantage of more accurately inferring its competitors' users, facilitating its marketing strategies and outcomes.

TABLE 4.4

THE NUMBER OF ACTIVE CALL USERS ACROSS OPERATORS

| | $O_0 \to O_0$ | $O_0 \to O_1$ | $O_0 \to O_2$ | $O_1 \to O_1$ | $O_1 \to O_0$ | $O_1 \to O_2$ | $O_2 \to O_2$ | $O_2 \to O_0$ | $O_2 \to O_1$ |
|---|---|---|---|---|---|---|---|---|---|
| #users | 608,589 | 608,589 | 608,589 | 292,848 | 292,848 | 292,848 | 183,893 | 183,893 | 183,893 |
| #edges | 1,291,086 | 534,064 | 342,845 | 424,394 | 534,064 | 205,487 | 208,452 | 342,845 | 205,487 |
| degree | 2.12 | 0.88 | 0.56 | 1.45 | 1.82 | 0.70 | 1.13 | 1.86 | 1.12 |

TABLE 4.5

THE NUMBER OF ACTIVE SMS USERS ACROSS OPERATORS

| | $O_0 \to O_0$ | $O_0 \to O_1$ | $O_0 \to O_2$ | $O_1 \to O_1$ | $O_1 \to O_0$ | $O_1 \to O_2$ | $O_2 \to O_2$ | $O_2 \to O_0$ | $O_2 \to O_1$ |
|---|---|---|---|---|---|---|---|---|---|
| #users | 161,547 | 161,547 | 161,547 | 87,556 | 87,556 | 87,556 | 56,634 | 56,634 | 56,634 |
| #edges | 257,154 | 123,192 | 72,313 | 93,342 | 123,192 | 46,807 | 37,660 | 72,313 | 46,807 |
| degree | 1.59 | 0.76 | 0.45 | 1.06 | 1.41 | 0.53 | 0.66 | 1.28 | 0.83 |

TABLE 4.6

COUPLED NETWORK DEMOGRAPHIC PREDICTION

| Network | Case | Gender | | | Age | | |
|---|---|---|---|---|---|---|---|
| | | wPrecision | wRecall | wF1-Measure | wPrecision | wRecall | wF1-Measure |
| CALL | $O_0$ to $O_1$ | 0.7870 | 0.7800 | 0.7807 | 0.7075 | 0.7087 | 0.7039 |
| | $O_0$ to $O_2$ | 0.7936 | 0.7939 | 0.7818 | 0.7100 | 0.7140 | 0.7085 |
| | $O_1$ to $O_0$ | 0.7404 | 0.7403 | 0.7396 | 0.6986 | 0.6801 | 0.6696 |
| | $O_1$ to $O_2$ | 0.7986 | 0.7979 | 0.7982 | 0.7160 | 0.7167 | 0.7094 |
| | $O_2$ to $O_0$ | 0.7325 | 0.7282 | 0.7251 | 0.6900 | 0.6758 | 0.6622 |
| | $O_2$ to $O_1$ | 0.7810 | 0.7794 | 0.7768 | 0.7147 | 0.7090 | 0.6981 |
| SMS | $O_0$ to $O_1$ | 0.7217 | 0.7222 | 0.7219 | 0.7172 | 0.7168 | 0.7049 |
| | $O_0$ to $O_2$ | 0.7329 | 0.7326 | 0.7327 | 0.7240 | 0.7259 | 0.7143 |
| | $O_1$ to $O_0$ | 0.6737 | 0.6713 | 0.6721 | 0.6897 | 0.6734 | 0.6540 |
| | $O_1$ to $O_2$ | 0.7347 | 0.7288 | 0.7285 | 0.7272 | 0.7245 | 0.7095 |
| | $O_2$ to $O_0$ | 0.6831 | 0.6846 | 0.6798 | 0.6885 | 0.6729 | 0.6497 |
| | $O_2$ to $O_1$ | 0.7232 | 0.7201 | 0.7143 | 0.7191 | 0.7152 | 0.6964 |

## 4.6 Related Work

The availability of mobile phone communication records has offered researchers many ways to analyze mobile networks, greatly enhancing our understanding of human mobile behavior [20, 44, 179].

To better model the macro properties of mobile communication networks, Onnela et al. [164] examine the local and global structure of a society-wide mobile communication network. Hidalgo and Rodriguez-Sickert [89] investigate the communication persistence in mobile phone networks. Faloutsos et al. [180] first propose the double pareto-lognormal distribution to model the macro properties in call networks, which is beyond power-law and lognormal distributions. They further discover that not only the node properties but also clique structures follow the power-law distribution in mobile networks [54]. Recently, the emergence of work on human mobility [47, 77, 223, 242] and mobile communication networks [7, 68, 194], where human activities are tracked by mobile phones, provides us a means of understanding and predicting mobile social behavior. Eagle et al. [55] try to infer the friendship network in mobile phone data. Tseng et al. [187] aim to discover the valuable user behavior patterns by mining in mobile commerce environments. Miritello et al. [149] discover that people follow underpinning strategies to interact with each other due to limited communication capacity. Meng et al. [143] study the correlations and differences between mobile and online networking behavior. Calabrese and Blondel et al. [20, 26] survey the problems, techniques, and results by using mobile phones network data. However, most previous work focuses on scaling the macroscopic properties of mobile networks, while our work incorporates the micro-network structure to model human communication behavior in mobile networks.

Furthermore, there are several works on user demographic and profile modeling. Existing works try to infer user demographics based on their online browsing [95], gaming [201] and search [17] behaviors. Herring surveys how online communications

facilitate gender equality, in particular, empowering women to achieve social identity that are difficult in offline environment [88]. Leskovec and Horvitz [120] examine the interplay of the MSN network and user demographic attributes. Mislove et al. study the demographics of Twitter users [151]. Tang et al. extract and model the researcher profiles in large-scale collaboration networks [203]. Matthew and Macskassy [144] analyze both the text and the network connectivity of the blogs to infer the demographics of bloggers. Dong et al. [43] investigate the mobile call duration behavior in mobile social networks and find that young females tend to make long phone calls [190], in particular in the evening. Llimona et al. [128] study the impact of gender and call duration on self-reported customer satisfaction. Macskassy et al. [29] also learn a label propagation model to infer users' public profiles in Facebook social network. Additionally, researchers have used network information to identify user status differences in email [48, 96] and LinkedIn networks [241]. Nokia research organized the 2012 Mobile Data Challenge to infer mobile user demographics by using 200 individual communication records without network information [152, 236]. Kovanen et al. [110] utilize temporal motifs to reveal demographic homophily in dynamic communication networks. The main difference between existing work and our efforts lies in that existing work mainly analyzes demographics (gender, age, status, etc.) separately, while our analysis and model consider the interrelation among different demographic attributes.

## 4.7 Conclusion

In this chapter, we model users' social decisions on connecting and maintaining relationships conditioned on their demographic profiles in large-scale mobile communication networks. We engage in answering the question of to what extent user demographics can be revealed from mobile communication interactions. We formalize a demographic prediction problem to simultaneously infer users' gender and age,

and further propose the WhoAmI method to solve it. Experimental results in phone call and text messaging networks demonstrate both the effectiveness and efficiency of our proposed model. Meanwhile, we identify a new problem—coupled network demographic prediction across multiple mobile operators. To address the unique challenges in this task, we present a coupled variant of the WhoAmI method. Our results unveil the predictability of user demographics across competitor networks, enabling the real-world application scenario of business intelligence in telecommunication.

Despite the promising predictive performance of the present method, there is still large room left for future work. First, in addition to model phone calls and text messages separately, it would be interesting to predict user demographics from the mobile network as a whole by combining the phone call and text messaging networks into one network. Second, some other social strategies and theories can be explored and validated for inferring user social traits and attributes. Finally, examining how the inferred demographics can help other topics in social network analysis, such as influence propagation, community detection, and network evolution, would also be very meaningful.

# PART II

# DIVERSITY IN BIG NETWORKS

# CHAPTER 5

# STRUCTURAL DIVERSITY AND EMBEDDEDNESS

## 5.1 Overview

Understanding the ways in which local network structures are formed and organized is a fundamental problem in network science. A widely recognized organizing principle of networks is structural homophily, which suggests that people with more common neighbors are more likely to connect with each other. However, what influence the diverse structures embedded in common neighbors (e.g.,  and  ) have on link formation is much less well understood. To explore this problem, in this chapter we begin by characterizing the structure of common neighborhoods as a function of their diversity and embeddedness. Using a collection of 120 large-scale networks—the biggest with over 60 million nodes and 1.8 billion edges—we then leverage these structural characteristics to develop a unique network signature, which we use to uncover several distinct network superfamilies not discoverable by conventional methods. We demonstrate that the impact of the common neighbor subgraph on link existence can vary substantially across networks, and we discover striking cases where it violates the principle of homophily. Our findings suggest that the common neighborhood signature (CNS) is an intrinsic network property, pointing to potential advancement in theories of network organization and evolution.

This chapter is largely extracted from a pre-print manuscript [51]. It is a joint work with Reid A. Johnson, Jian Xu, and Nitesh V. Chawla.

## 5.2 Introduction

Since the time of Aristotle, it's been known that people "love those who are like themselves" [10]. We now know this as the principle of homophily, which suggests that the tendency of individuals to associate and bond with similar others drives the formation of social relationships [115, 140]. The powerful effects of homophily pervade our everyday lives, silently influencing our most basic relationships from friendship to marriage [140]. By guiding the formation of relationships, homophily also plays an important role in the dissemination of information, behavior, and even health [33]. But homophily applies to more than just shared traits or characteristics: it applies to the fundamental structure of our relationships as well.

Structural homophily holds that individuals with more friends in common are more likely to associate [99, 161]. The tendency of individuals to connect based on structural homophily has been widely explored in network science [3, 125], where it has been shown to be a strong driving force of link formation over a large assortment of networks. Yet, while structural homophily accounts for similarity based on the actual number of common neighbors, it fails to account for the diverse ways in which these neighbors may be embedded—instead accounted for by phenomena known as structural diversity [216] and embeddedness [81]. Despite the well-studied importance of structural homophily, the effects of diversity and embeddedness are much less well-understood. This leaves many interesting questions concerning the role of the structure of common neighborhoods unanswered, including how this structure manifests across networks, how it varies according to the type of network, how well it concords with the principle of homophily, and how it affects network connectivity in a neighborhood and beyond.

**Motivating example.** Consider the real-world scenarios presented in Figure 5.1. According to structural homophily, the probability that two users $v_i$ and $v_j$ know each

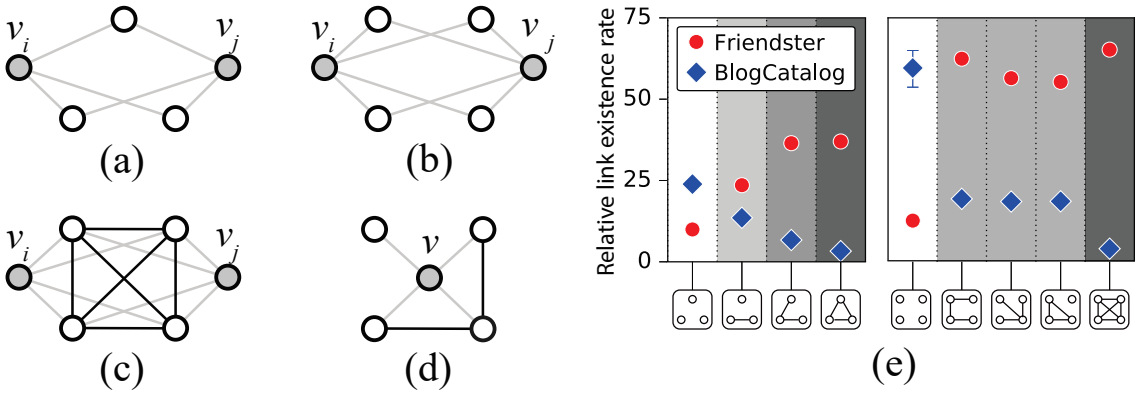



Figure 5.1. Structural diversity and embeddedness of common neighborhoods. Two nodes $v_i$ and $v_j$ with three disconnected (a), four disconnected (b), and four connected (c) common neighbors. (d) Different from structural diversity in an ego-centric notion [216], we go beyond an ego, and focus on the structural diversity and embeddedness of common neighborhoods for two persons. (e) Structural diversity and embeddedness of common neighborhoods affect link existence rate.

other given that they share four common neighbors (CN), as shown in Figure 5.1(b), is generally higher than when they share only three, as shown in Figure 5.1(a). Formally, $P(e_{ij}{=}1|\#\text{CN}{=}4) > P(e_{ij}{=}1|\#\text{CN}{=}3)$, where $e_{ij}{=}1$ denotes the existence of an edge $e$ between $v_i$ and $v_j$. A natural question that arises is how two users' common neighborhood—that is, the subgraph structure of their common neighbors—influences the probability that they form a link in the network. For example, let us assume that $v_i$ and $v_j$ share four common neighbors. Are $v_i$ and $v_j$ more likely to connect with each other if their four common neighbors do not know each other (i.e., ⚀), as in Figure 5.1(b), or if they all know each other (i.e., ⚅), as in Figure 5.1(c)? In essence, then, we are interested in the truth of the following inequality:

$$P(e_{ij}{=}1 \mid ⚀) \gtrless P(e_{ij}{=}1 \mid ⚅) \ ?$$

In this chapter, we formally define the structure of common neighborhoods between two individuals and study its impact on the probability that these individuals

know each other. Our definition of the structure of a common neighborhood is a mixture of the diversity and embeddedness of common social contexts. Diversity measures the number of connected components that comprise the common neighborhood, capturing the variable ways in which a neighborhood may be composed [216]. Embeddedness measures the ratio of the number of actual edges to the number of all possible edges among common neighbors, capturing "the extent that a dyad's mutual contacts are connected to one another" [79, 81, 219]. Using this formulation, we study the influence of structural diversity on the formation of (social) relationships across more than one hundred large-scale networks from a wide range of domains (cf. Section 5.3), *making this the largest empirical analysis done on networks to date.*

We leverage our definition of the structural diversity and embeddedness of common neighborhoods to develop a signature for network superfamily detection. By employing the common neighborhood signature (CNS) of each network, we are able to cluster all of the real-world networks used in our study into three distinct superfamilies, each of which displays unique link existence patterns. We find that these superfamilies cannot be uncovered by subgraph significance profiles [148] and other properties, indicating that the common neighborhood signature provides the ability to unveil previously undiscovered mechanisms of network organization. This finding demonstrates that the structure of common neighborhoods can effectively capture driving forces intrinsic to the formation of local network structures.

We examine how the structure of common neighborhoods impacts link existence and network connectivity, with in-depth investigations into three large-scale networks—Friendster, BlogCatalog, and YouTube. Each network is representative of a distinct network superfamily, with findings that generalize to all networks within the superfamily. Figure 5.1(e) reports the relative link existence rate between each pair of users who have at least one common neighbor, conditioned on several representative common neighborhoods ($x$-axis). We observe that with the same edge

density (embeddedness), the link existence rates given different common neighborhoods (e.g.,  ,  , and  ) are relatively close. However, we find that when we fix the size of the common neighborhood (e.g., three common neighbors), an increase in the structural diversity of the neighborhood (i.e., #components) negatively impacts the formation of online friendships in Friendster and its network superfamily—that is, $P(e{=}1|$  $) < P(e{=}1|$  $)$—while it actually facilitates link formation in BlogCatalog and its superfamily—that is, $P(e{=}1|$  $) > P(e{=}1|$  $)$. Other network properties—including degree distribution, degree sequence, and subgraph frequency—show strong similarities across the different superfamilies, and thus cannot adequately characterize the differences we observe in the structural diversity of common neighborhoods.

We also discover striking phenomena where structural diversity and embeddedness violate the principle of homophily. When applied to the context of common neighborhoods, the principle suggests that $P(e{=}1 \mid \#\text{CN}{=}4) > P(e{=}1 \mid \#\text{CN}{=}3)$. However, if we consider BlogCatalog, for example, we find that the link existence rate of four common neighbors in a single component is significantly lower than the rate of only three disconnected common neighbors, i.e., $P(e{=}1|$  $) < P(e{=}1|$  $)$. Similarly, in Friendster, homophily is violated when comparing four disconnected common neighbors with three connected common neighbors, i.e., $P(e{=}1|$  $) > P(e{=}1|$  $)$.

Studying structural diversity and embeddedness in the context of common neighborhoods sheds light on the pursuit to understand the driving forces behind the organization of neighborhoods in social networks. Our findings also have important, practical implications for recommendation functions in social networks, such as "People You May Know (PYMK)" in Facebook and LinedIn, as well as "Who to Follow" in Twitter.

## 5.3 Big Network Data

To comprehensively examine our proposed concept of the structural diversity of common neighborhoods, we have assembled a large collection of big network datasets from several well-known data platforms, including (in alphabetical order): AMiner (AMiner Open Science Platform [203]), ASU (Social Computing Data Repository [238]), KONECT (Koblenz Network Collection [114]), MPI (Social Computing Research at MPI-SWS [150]), ND (Notre Dame [14]), NetRep (Network Data Repository [177]), Newman [162], and SNAP (Stanford Large Network Dataset Collection [191]).

In total, we have compiled a set of 120 large-scale undirected and unweighted networks from the platforms listed above, including 80 real-world networks and 40 random graphs. We have cleaned the networks as follows: For directed social networks, such as mobile phone and SMS networks, we retain only reciprocal connections as undirected edges. For other directed networks that have no reciprocal connection, such as citation networks, we convert each directed link into an undirected one. We have then pruned the resulting undirected networks by removing all duplicate edges and self-loops, retaining only the largest connected component.

The 80 real networks used in this chapter are shown in Table 5.1, in which *cc* denotes clustering coefficient and *d* denotes diameter. Due to the large set of networks, we have labeled them according to the follow nomenclature: *type-original-platform*. *type* denotes the network type, most of which have been previously designated by their source data platforms (see the taxonomy example in SNAP). For a given network, *type* can be one of social blog-based (*blog-*), collaboration (*ca-*), citation (*cit-*), communication (*comm-*), location-based (*loc-*), online social (*soc-*) networks, or web hyperlink graphs (*web-*). *original* denotes the original name of the networks as provided by each platform. $platform$ denotes the data platform from which the network has been sourced. We note that the largest network used in this work is the

*soc-Friendster-SNAP* online social network, which consists of over 65 million nodes and more than 1.8 billion edges.

To ensure that our study explores a representative sample of network structures, we also test the concept of structural diversity and embeddedness of common neighborhoods on 40 random graphs, including 10 networks generated by each of the four models: the Erdős-Rényi (ER) model [59], Barabási-Albert (BA) model [14], Watts-Strogatz (WS) model [228], and Kronecker model [123]. For the first three models, the number of nodes is set as 1,000,000. In the ER model, we set the edge creation probability to between $5\times10^{-6}$ and $5\times10^{-5}$ with a step of $5\times10^{-6}$, thereby generating 10 ER random graphs with the number of edges ranging from roughly 2,000,000 to 25,000,000. We use the BA model to generate 10 BA random graphs with between 2,000,000 and 20,000,000 edges. We use the WS model to generate 10 WS random graphs by setting different mean degrees $k$ and rewiring probabilities $\beta$, where $k$ is chosen from 8, 12, 16, 20, and 24, and $\beta$ is 0.2 or 0.8. There are between 2,000,000 and 14,000,000 edges in the WS graphs. Finally, we use 10 Kronecker graphs with the original parameters (Estimated by Leskovec et al. [123]). fitted to 10 real networks, which are among the 80 we use above. That means, we would expect the results discovered from the 10 Kronecker graphs be in close agreement with the corresponding 10 real networks, if Kronecker model was also capable of preserving the common neighborhood in addition to traditional network properties, such as degree, triangle, and diameter distributions.

TABLE 5.1

THE STATISTICS OF 120 NETWORKS

| Network Name | #nodes | #edges | degree | average cc | global cc | $d$ | #triangles | #triplets |
|---|---|---|---|---|---|---|---|---|
| blog-BlogCatalog1-ASU | 88784 | 2093195 | 47 | 0.3533 | 0.0624 | 9 | 51193389 | 2410854351 |
| blog-BlogCatalog2-ASU | 97884 | 1668647 | 34 | 0.4921 | 0.0403 | 5 | 40662527 | 2986356565 |
| blog-BlogCatalog3-ASU | 10312 | 333983 | 64 | 0.4632 | 0.0973 | 5 | 5608664 | 167281662 |
| ca-Actor-ND | 374511 | 15014839 | 80 | 0.7788 | 0.1867 | 13 | 346728049 | 5225759780 |
| ca-AstroPh-Newman | 14845 | 119652 | 16 | 0.6696 | 0.5937 | 14 | 754159 | 3056896 |
| ca-AstroPh-SNAP | 17903 | 196972 | 22 | 0.6328 | 0.4032 | 14 | 1350014 | 8694840 |
| ca-CondMat2003-Newman | 27519 | 116181 | 8 | 0.6546 | 0.3393 | 16 | 228093 | 1788720 |
| ca-CondMat2005-Newman | 36458 | 171735 | 9 | 0.6566 | 0.2903 | 18 | 374300 | 3493465 |
| ca-CondMat-SNAP | 21363 | 91286 | 8 | 0.6417 | 0.3172 | 15 | 171051 | 1446763 |
| ca-CS2004to2008-AMiner | 434357 | 1578275 | 7 | 0.6684 | 0.3705 | 27 | 3451794 | 24501202 |
| ca-CS2006to2010-AMiner | 543452 | 2066296 | 7 | 0.6745 | 0.2939 | 25 | 4372725 | 40256430 |
| ca-CS2009to2010-AMiner | 315263 | 1059740 | 6 | 0.6989 | 0.4181 | 23 | 2115515 | 13063018 |
| ca-CS2011to2012-Aminer | 347389 | 1229716 | 7 | 0.7073 | 0.4004 | 29 | 2585990 | 16787160 |

TABLE 5.1

*Continued*

| Network Name | #nodes | #edges | degree | average cc | global cc | $d$ | #triangles | #triplets |
|---|---|---|---|---|---|---|---|---|
| ca-CS-AMiner | 1066379 | 4594140 | 8 | 0.6496 | 0.1873 | 24 | 10312677 | 154825662 |
| ca-DBLP-SNAP | 317080 | 1049866 | 6 | 0.6324 | 0.3850 | 23 | 2224385 | 15107734 |
| ca-GrQc-SNAP | 4158 | 13422 | 6 | 0.5569 | 1.0829 | 17 | 47779 | 84582 |
| ca-HepPh-SNAP | 11204 | 117619 | 20 | 0.6216 | 1.1768 | 13 | 3357890 | 5202255 |
| ca-HepTh-SNAP | 8638 | 24806 | 5 | 0.4816 | 0.3460 | 18 | 27869 | 213790 |
| ca-Hollywood-NetRep | 1069126 | 56306653 | 105 | 0.7664 | 0.3900 | 12 | 4916220615 | 32896279137 |
| ca-MathSci-NetRep | 332689 | 820644 | 4 | 0.4104 | 0.1504 | 24 | 576778 | 10928378 |
| loc-Brightkite-SNAP | 56739 | 212945 | 7 | 0.1734 | 0.1193 | 18 | 494408 | 11938424 |
| loc-Foursquare-ASU | 639014 | 3214986 | 10 | 0.1080 | 0.0016 | 4 | 21651003 | 39400700856 |
| loc-FourSquare-NetRep | 639014 | 3214986 | 10 | 0.1080 | 0.0016 | 4 | 21651003 | 39400700856 |
| loc-Gowalla-SNAP | 196591 | 950327 | 9 | 0.2367 | 0.0239 | 16 | 2273138 | 283580626 |
| cit-CiteSeer-KONECT | 365154 | 1721981 | 9 | 0.1832 | 0.0513 | 34 | 1350310 | 77658938 |
| cit-Cora-KONECT | 23166 | 89157 | 7 | 0.2660 | 0.1268 | 20 | 78791 | 1786074 |

TABLE 5.1

*Continued*

| Network Name | #nodes | #edges | degree | average cc | global cc | $d$ | #triangles | #triplets |
|---|---|---|---|---|---|---|---|---|
| cit-HepPh-d-SNAP | 34401 | 420784 | 24 | 0.2856 | 0.1613 | 14 | 1276859 | 22468237 |
| cit-HepTh-d-SNAP | 27400 | 352021 | 25 | 0.3139 | 0.1299 | 15 | 1478698 | 32665296 |
| cit-Patents-d-SNAP | 3764117 | 16511740 | 8 | 0.0758 | 0.0703 | 26 | 7514922 | 313229094 |
| comm-CALL-ND | 4295638 | 7893769 | 3 | 0.2179 | 0.1985 | 45 | 2253963 | 31804482 |
| comm-EmailEnron-SNAP | 33696 | 180811 | 10 | 0.5092 | 0.0903 | 13 | 725311 | 23384268 |
| comm-EmailEuAll-SNAP | 32430 | 54397 | 3 | 0.1127 | 0.0273 | 9 | 48992 | 5341634 |
| comm-LinuxKernel-KONECT | 10857 | 76317 | 14 | 0.3486 | 0.1185 | 13 | 698240 | 16977912 |
| comm-Mobile-ND | 5324963 | 10410903 | 3 | 0.1811 | 0.0104 | 36 | 2895897 | 835604293 |
| comm-SMS-ND | 2369078 | 3330086 | 2 | 0.0669 | 0.0013 | 42 | 326282 | 770920401 |
| comm-WikiTalk-SNAP | 92117 | 360767 | 7 | 0.0589 | 0.0483 | 11 | 836467 | 51083880 |
| web-BaiduBaike-KONECT | 2107689 | 16996139 | 16 | 0.1171 | 0.0025 | 20 | 25206270 | 30809207121 |
| web-BerkStan-SNAP | 654782 | 6581871 | 20 | 0.6066 | 0.0069 | 208 | 64520617 | 27786200608 |
| web-Google-SNAP | 855802 | 4291352 | 10 | 0.5190 | 0.0572 | 24 | 13356298 | 686679376 |

TABLE 5.1

*Continued*

| Network Name | #nodes | #edges | degree | average cc | global cc | $d$ | #triangles | #triplets |
|---|---|---|---|---|---|---|---|---|
| web-Hudong-KONECT | 1962418 | 14419760 | 14 | 0.0783 | 0.0035 | 16 | 21611635 | 18660819412 |
| web-Stanford-SNAP | 255265 | 1941926 | 15 | 0.6189 | 0.0086 | 164 | 11277977 | 3907779392 |
| web-WWW-ND | 325729 | 1090108 | 6 | 0.2346 | 0.0931 | 46 | 8910005 | 278151159 |
| soc-Academia-NetRep | 137969 | 369692 | 5 | 0.1421 | 0.0806 | 21 | 220641 | 7995452 |
| soc-Advogato-KONECT | 2716 | 7773 | 5 | 0.2233 | 0.1325 | 13 | 5383 | 116510 |
| soc-BuzzNet-ASU | 101163 | 2763066 | 54 | 0.2321 | 0.0108 | 5 | 30919848 | 8542533935 |
| soc-Catster-NetRep | 148826 | 5447464 | 73 | 0.3877 | 0.0111 | 10 | 185462078 | 50059386906 |
| soc-Delicious-ASU | 536108 | 1365961 | 5 | 0.0322 | 0.0106 | 14 | 487972 | 137770815 |
| soc-Digg-ASU | 770799 | 5907132 | 15 | 0.0881 | 0.0482 | 18 | 62710792 | 3842962151 |
| soc-Dogster-NetRep | 426485 | 8543321 | 40 | 0.1710 | 0.0144 | 11 | 83499345 | 17303939974 |
| soc-Douban-ASU | 154908 | 327162 | 4 | 0.0161 | 0.0104 | 9 | 40612 | 11623280 |
| soc-Epinions1-d-SNAP | 75877 | 405739 | 10 | 0.1378 | 0.0687 | 15 | 1624481 | 69327677 |
| soc-Facebook-MPI | 63392 | 816886 | 25 | 0.2218 | 0.1639 | 15 | 3501534 | 60606675 |
| soc-Flickr-AMiner | 214424 | 9114421 | 85 | 0.1464 | 0.0832 | 10 | 132139697 | 4630544599 |

TABLE 5.1

*Continued*

| Network Name | #nodes | #edges | degree | average cc | global cc | $d$ | #triangles | #triplets |
|---|---|---|---|---|---|---|---|---|
| soc-Flickr-ASU | 80513 | 5899882 | 146 | 0.1652 | 0.2142 | 6 | 271601126 | 3531448904 |
| soc-Flickr-MPI | 1624992 | 15476835 | 19 | 0.1892 | 0.1212 | 24 | 548646525 | 13028541364 |
| soc-Flixter-ASU | 2523386 | 7918801 | 6 | 0.0834 | 0.0138 | 8 | 7897122 | 1711880027 |
| soc-Friendster-ASU | 5689498 | 14067887 | 4 | 0.0502 | 0.0048 | 9 | 8722131 | 5484816732 |
| soc-Friendster-SNAP | 65608366 | 1806067135 | 55 | 0.1623 | 0.0176 | 37 | 4173724142 | 708133792538 |
| soc-GooglePlus-NetRep | 78723 | 319999 | 8 | 0.1982 | 0.2934 | 59 | 1386340 | 12787287 |
| soc-Hamsterster-KONECT | 2000 | 16098 | 16 | 0.5401 | 0.2709 | 10 | 52665 | 530614 |
| soc-Hyves-ASU | 1402673 | 2777419 | 3 | 0.0448 | 0.0016 | 10 | 752401 | 1444870827 |
| soc-LastFM-AMiner | 135876 | 1685158 | 24 | 0.1983 | 0.0946 | 12 | 9097399 | 279291397 |
| soc-LastFM-ASU | 1191805 | 4519330 | 7 | 0.0727 | 0.0131 | 10 | 3946207 | 898270114 |
| soc-Libimseti-KONECT | 34339 | 124722 | 7 | 0.0224 | 0.0265 | 15 | 54375 | 6103992 |
| soc-LinkedIn-AMiner | 6725712 | 19360071 | 5 | 0.3700 | 0.2863 | 32 | 12862009 | 121917817 |
| soc-LiveJournal1-d-SNAP | 4843953 | 42845684 | 17 | 0.2743 | 0.1280 | 20 | 285688896 | 6412296576 |
| soc-LiveJournal-AMiner | 3017282 | 85654975 | 56 | 0.1196 | 0.0017 | 8 | 507338233 | 919635317380 |

TABLE 5.1

*Continued*

| Network Name | #nodes | #edges | degree | average cc | global cc | $d$ | #triangles | #triplets |
|---|---|---|---|---|---|---|---|---|
| soc-LiveJournal-ASU | 2238731 | 12816184 | 11 | 0.1270 | 0.0230 | 8 | 28204049 | 3658174479 |
| soc-LiveJournal-MPI | 5189809 | 48688097 | 18 | 0.2749 | 0.1352 | 23 | 310784143 | 6586074658 |
| soc-LiveJournal-SNAP | 3997962 | 34681189 | 17 | 0.2843 | 0.1368 | 21 | 177820130 | 3722307805 |
| soc-LiveMocha-ASU | 104103 | 2193083 | 42 | 0.0544 | 0.0142 | 6 | 3361651 | 706231197 |
| soc-MySpace-AMiner | 853360 | 5635236 | 13 | 0.0433 | 0.0022 | 14 | 1256533 | 1686861075 |
| soc-Orkut-NetRep | 2997166 | 106349209 | 70 | 0.1700 | 0.0439 | 9 | 524643952 | 35294034217 |
| soc-Orkut-SNAP | 3072441 | 117185083 | 76 | 0.1666 | 0.0424 | 9 | 627584181 | 43742714028 |
| soc-Pokec-d-SNAP | 1632803 | 22301964 | 27 | 0.1094 | 0.0483 | 14 | 32557458 | 1988401184 |
| soc-Prosper-d-KONECT | 89171 | 3329970 | 74 | 0.0049 | 0.0031 | 8 | 1158669 | 1108949447 |
| soc-Slashdot0811-d-SNAP | 77360 | 469180 | 12 | 0.0555 | 0.0246 | 12 | 551724 | 66861129 |
| soc-Slashdot0902-d-SNAP | 82168 | 504230 | 12 | 0.0603 | 0.0245 | 13 | 602592 | 73175813 |
| soc-WikiVote-d-SNAP | 7066 | 100736 | 28 | 0.1419 | 0.1369 | 7 | 608389 | 12720410 |
| soc-YouTube-MPI | 1134890 | 2987624 | 5 | 0.0808 | 0.0062 | 24 | 3056386 | 1465313402 |
| random-ba-2 | 1000000 | 1999996 | 3 | 0.0001 | 0.0000 | 11 | 435 | 40304220 |

TABLE 5.1

*Continued*

| Network Name | #nodes | #edges | degree | average cc | global cc | $d$ | #triangles | #triplets |
|---|---|---|---|---|---|---|---|---|
| random-ba-4 | 1000000 | 3999984 | 7 | 0.0001 | 0.0001 | 8 | 3933 | 137191058 |
| random-ba-6 | 1000000 | 5999964 | 11 | 0.0002 | 0.0001 | 7 | 11293 | 270899331 |
| random-ba-8 | 1000000 | 7999936 | 15 | 0.0002 | 0.0002 | 6 | 27328 | 479626736 |
| random-ba-10 | 1000000 | 9999900 | 19 | 0.0003 | 0.0002 | 6 | 51315 | 718878801 |
| random-ba-12 | 1000000 | 11999856 | 23 | 0.0003 | 0.0003 | 5 | 84034 | 1004827944 |
| random-ba-14 | 1000000 | 13999804 | 27 | 0.0003 | 0.0003 | 5 | 129967 | 1344161988 |
| random-ba-16 | 1000000 | 15999744 | 31 | 0.0004 | 0.0003 | 5 | 187591 | 1717526329 |
| random-ba-18 | 1000000 | 17999676 | 35 | 0.0004 | 0.0004 | 5 | 268499 | 2188921333 |
| random-ba-20 | 1000000 | 19999600 | 39 | 0.0004 | 0.0004 | 5 | 352033 | 2631857430 |
| random-er-5e-06 | 993242 | 2498898 | 5 | 0.0000 | 0.0000 | 16 | 20 | 12484231 |
| random-er-1e-05 | 999945 | 5001782 | 10 | 0.0000 | 0.0000 | 10 | 166 | 50052665 |
| random-er-1.5e-05 | 1000000 | 7504667 | 15 | 0.0000 | 0.0000 | 8 | 619 | 112634285 |
| random-er-2e-05 | 1000000 | 9998180 | 19 | 0.0000 | 0.0000 | 7 | 1366 | 199917141 |
| random-er-2.5e-05 | 1000000 | 12504290 | 25 | 0.0000 | 0.0000 | 6 | 2630 | 312702098 |

TABLE 5.1

*Continued*

| Network Name | #nodes | #edges | degree | average cc | global cc | $d$ | #triangles | #triplets |
|---|---|---|---|---|---|---|---|---|
| random-er-3e-05 | 1000000 | 14998486 | 29 | 0.0000 | 0.0000 | 6 | 4530 | 449858883 |
| random-er-3.5e-05 | 1000000 | 17498938 | 34 | 0.0000 | 0.0000 | 5 | 7274 | 612427121 |
| random-er-4e-05 | 1000000 | 20002664 | 40 | 0.0000 | 0.0000 | 5 | 10729 | 800174659 |
| random-er-4.5e-05 | 1000000 | 22499181 | 44 | 0.0000 | 0.0000 | 5 | 15196 | 1012356360 |
| random-er-5e-05 | 1000000 | 24997692 | 49 | 0.0000 | 0.0000 | 5 | 20825 | 1249733430 |
| random-k20-800-600-600-400 | 1048287 | 20099931 | 38 | 0.0001 | 0.0001 | 7 | 47331 | 1332866503 |
| random-k20-847-641-641-072 | 740760 | 3559350 | 9 | 0.0004 | 0.0001 | 10 | 11343 | 249748200 |
| random-k20-900-400-400-600 | 1047241 | 8579924 | 16 | 0.0002 | 0.0003 | 11 | 19917 | 196561115 |
| random-k20-900-600-600-100 | 717728 | 3527146 | 9 | 0.0006 | 0.0003 | 12 | 24180 | 284289856 |
| random-k20-954-594-594-019 | 523915 | 2466409 | 9 | 0.0020 | 0.0005 | 10 | 62042 | 357289819 |
| random-k20-987-571-571-049 | 554986 | 2884877 | 10 | 0.0028 | 0.0008 | 10 | 131738 | 473991798 |
| random-k20-999-245-245-691 | 1003562 | 2929395 | 5 | 0.0052 | 0.0071 | 20 | 56717 | 24064578 |
| random-k20-999-271-271-587 | 870759 | 1810525 | 4 | 0.0019 | 0.0038 | 23 | 16253 | 12983663 |
| random-k20-999-307-307-574 | 962241 | 3130762 | 6 | 0.0010 | 0.0020 | 18 | 25714 | 39090320 |

TABLE 5.1

*Continued*

| Network Name | #nodes | #edges | degree | average cc | global cc | $d$ | #triangles | #triplets |
|---|---|---|---|---|---|---|---|---|
| random-k20-999-437-437-484 | 1040130 | 14000384 | 26 | 0.0002 | 0.0005 | 11 | 169897 | 948885915 |
| random-ws-8-0.2 | 1000000 | 4000000 | 8 | 0.3346 | 0.4081 | 12 | 3071207 | 19507015 |
| random-ws-8-0.8 | 1000000 | 4000000 | 8 | 0.0051 | 0.0048 | 9 | 48121 | 29769537 |
| random-ws-12-0.2 | 1000000 | 6000000 | 12 | 0.3532 | 0.4459 | 10 | 7684734 | 44022926 |
| random-ws-12-0.8 | 1000000 | 6000000 | 12 | 0.0054 | 0.0053 | 7 | 120214 | 68515178 |
| random-ws-16-0.2 | 1000000 | 8000000 | 16 | 0.3619 | 0.4644 | 8 | 14354463 | 78375140 |
| random-ws-16-0.8 | 1000000 | 8000000 | 16 | 0.0056 | 0.0055 | 7 | 225119 | 123159468 |
| random-ws-20-0.2 | 1000000 | 10000000 | 20 | 0.3660 | 0.4739 | 7 | 23023926 | 122732295 |
| random-ws-20-0.8 | 1000000 | 10000000 | 20 | 0.0056 | 0.0056 | 6 | 359942 | 193720514 |
| random-ws-24-0.2 | 1000000 | 12000000 | 24 | 0.3694 | 0.4814 | 7 | 33788586 | 176784043 |
| random-ws-24-0.8 | 1000000 | 12000000 | 24 | 0.0057 | 0.0057 | 6 | 530929 | 280173038 |

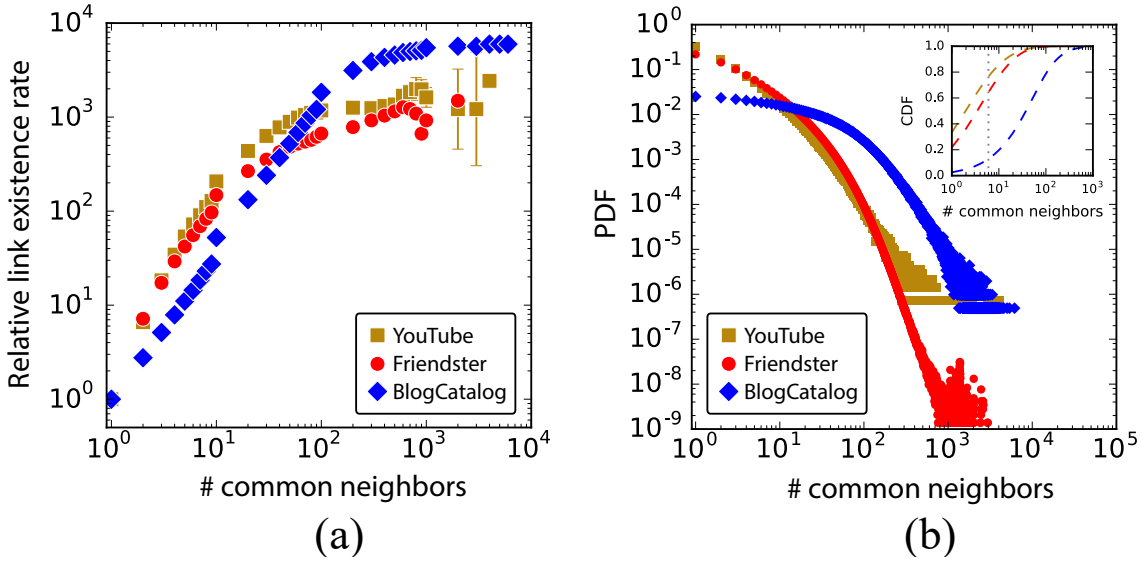



Figure 5.2. Common neighbor characterization. (a) The link existence rate as the function of #common neighbors. (b) The probability density function (PDF) of #common neighbors.

## 5.4 Common Neighborhood Signature (CNS)

In this section, we investigate the principles that drive the formation and organization of local structures in large-scale networks. In particular, we focus on the common neighborhood of each pair of users and ask the following question: How does a pair of individuals' common neighborhood—the subgraph with their mutual neighbors as nodes and the connections among them as edges—influence the probability that there exists a link between them?

Formally, we use $G = (V, E)$ to denote an undirected and unweighted network, where $V = \{v_i\}$ represents the set of nodes and $E \subseteq V \times V$ represents the set of links between two nodes. We denote each existing link, $e_{ij} \in E$, as $e_{ij} = (v_i, v_j) = 1$ and each non-existing link, $e_{ij} \notin E$, as $e_{ij} = (v_i, v_j) = 0$.

**Definition 2** ***Common Neighborhood:*** *Let $N(v_i)$ denote the adjacency list of a node $v_i$, i.e., $v_i$'s neighborhood. The common neighborhood of each pair of two nodes $v_i$ and $v_j$ can be represented as the subgraph composed of their common neighbors,*

*$G^{ij} = (V^{ij}, E^{ij})$, where $V^{ij} = N(v_i) \cap N(v_j)$ denotes the common neighbors of $v_i$ and $v_j$, and $E^{ij} = \{e_{pq} \mid e_{pq} \in E, v_p \in V^{ij}, v_q \in V^{ij}\}$ denotes the edges among their common neighbors.*

**Input:** Given a network $G = (V, E)$, the input of our problem includes 1) each pair of users who have at least one common neighbor, i.e., $\{(v_i, v_j) = e_{ij} \mid |V^{ij}| \geq 1\}$ , and the common neighborhood $G^{ij} = (V^{ij}, E^{ij})$ of each pair of users $v_i$ and $v_j$.

**Structural homophily.** The principle of structural homophily suggests that with more common neighbors, it is more likely for two people to know each other. Formally, this means that if $y > x$, then $P(e_{ij}{=}1 \mid |V^{ij}|{=}y)$ should generally be larger than $P(e_{ij}{=}1 \mid |V^{ij}|{=}x)$. A long line of work from various fields has demonstrated that this principle holds across a wide variety of different networks. For example, Figure 5.2(a) reports the link existence rate between two users ($y$-axis), conditioned on the size of their common neighborhood ($x$-axis) in three representative networks. We can see that as the number of common neighbors increases, the probability that two users are connected with each other increases in all three networks as well.

In this study, we revisit this principle of structural homophily, further proposing to study the structure—specifically, the diversity and embeddedness—of common neighborhoods. We characterize the structure of a small graph as a function of its diversity and embeddedness, formalizing its application to common neighborhoods.

**Definition 3** ***Structural Diversity and Embeddedness of Common Neighborhoods:*** *Given a network, $G$, a pair of users in this network, $v_i$ and $v_j$, and the pair's common neighborhood, $G^{ij} = (V^{ij}, E^{ij})$, we define the structure of $G^{ij}$ as a mixture of its diversity, $|C(G^{ij})|$, and embeddedness, $d(G^{ij})$, where $C(G^{ij})$ denotes the connected components in $G^{ij}$ and $d(G^{ij})$ denotes the edge density of $G^{ij}$.*

Consider a pair of users with four common neighbors. This pair's common neighborhood has 11 possible configuration structures:  ,  ,  ,  ,  ,  ,  ,

 ,  ,  , and  . In general, we refer to the more diverse structure as the one with fewer edges and more components.

**Output:** Our goal is to study the relations between the structure of two users' common neighborhood and the probability that there exists a link between these two users. Therefore, given two users $v_i$ and $v_j$ and their common neighborhood $G^{ij}$, the output of our problem is the link existence probability distribution of $G^{ij}$, that is, $P(e_{ij}{=}1 \mid G^{ij})$.

In each network, we enumerate all pairs of users who have at least one common neighbor. If we fix the common neighborhood $G^{ij}$ of two users $v_i$ and $v_j$, then we can compute the link existence probability $P(e_{ij}{=}1 \mid G^{ij})$ based on the number of link pairs that exist. To facilitate the comparability of results across networks with diverse sizes and densities, we define the relative link existence rate $R(e_{ij}{=}1 \mid G^{ij})$ as

$$R(e_{ij}{=}1 \mid G^{ij}) = \frac{P(e_{ij}{=}1 \mid G^{ij})}{P(e{=}1 \mid \#CN{=}1)},$$

where $P(e{=}1 \mid \#CN{=}1)$ denotes the link existence probability when two users have exactly one common neighbor.

**Definition 4** ***Common Neighborhood Signature (CNS):*** *Given a network $G = (V, E)$, its common neighborhood signature is defined as a vector of relative link existence rates with respect to the specified common neighborhoods.*

Consider, for example, user pairs with between two and four common neighbors. The common neighborhoods represented by these pairs of users correspond to a vector of the relative link existence rates for 17 subgraphs (2 subgraphs for common neighborhoods with size two, 4 for those with size three, and 11 for those with size four).

Given this input and output, our work seeks to understand the underlying driving forces behind link formation and network organization by answering the following

questions:

- Does the common neighborhood signature (CNS) vary across various of networks?
- Is the common neighborhood signature (CNS) a fundamental property of networks?
- How does the structural configuration of common neighborhoods influence the link existence probability?
- Does structural diversity and embeddedness concord or conflict with the principle of homophily in networks?
- Can structural diversity and embeddedness help to improve link inference?

## 5.5 CNS for Network Superfamilies

The global structure of network systems are governed by natural laws, through which constant, universal properties such as long-tailed degree distributions arise. However, even when subject to the global structures prescribed by these laws, different networks can still reveal distinct local properties and structures. One striking example is the discovery that networks with long-tailed degree distributions can be naturally cataloged into distinct superfamilies of networks based on their subgraph frequencies [148]. In this section, we investigate how the common neighborhood signature can—like subgraph frequency—uncover previously undiscovered mechanisms of network organization, thereby allowing it to serve as a fundamental property by which to catalog networks.

### 5.5.1 Network Superfamilies

To answer this question, we examine the similarity between the functions of structural diversity of common neighborhoods across the 120 networks studied. We begin by constructing, for each network, the common neighborhood signature represented by neighborhoods with between two and four common neighbors (constituting a 17-length vector). We then use this signature to compute the similarity in the structural

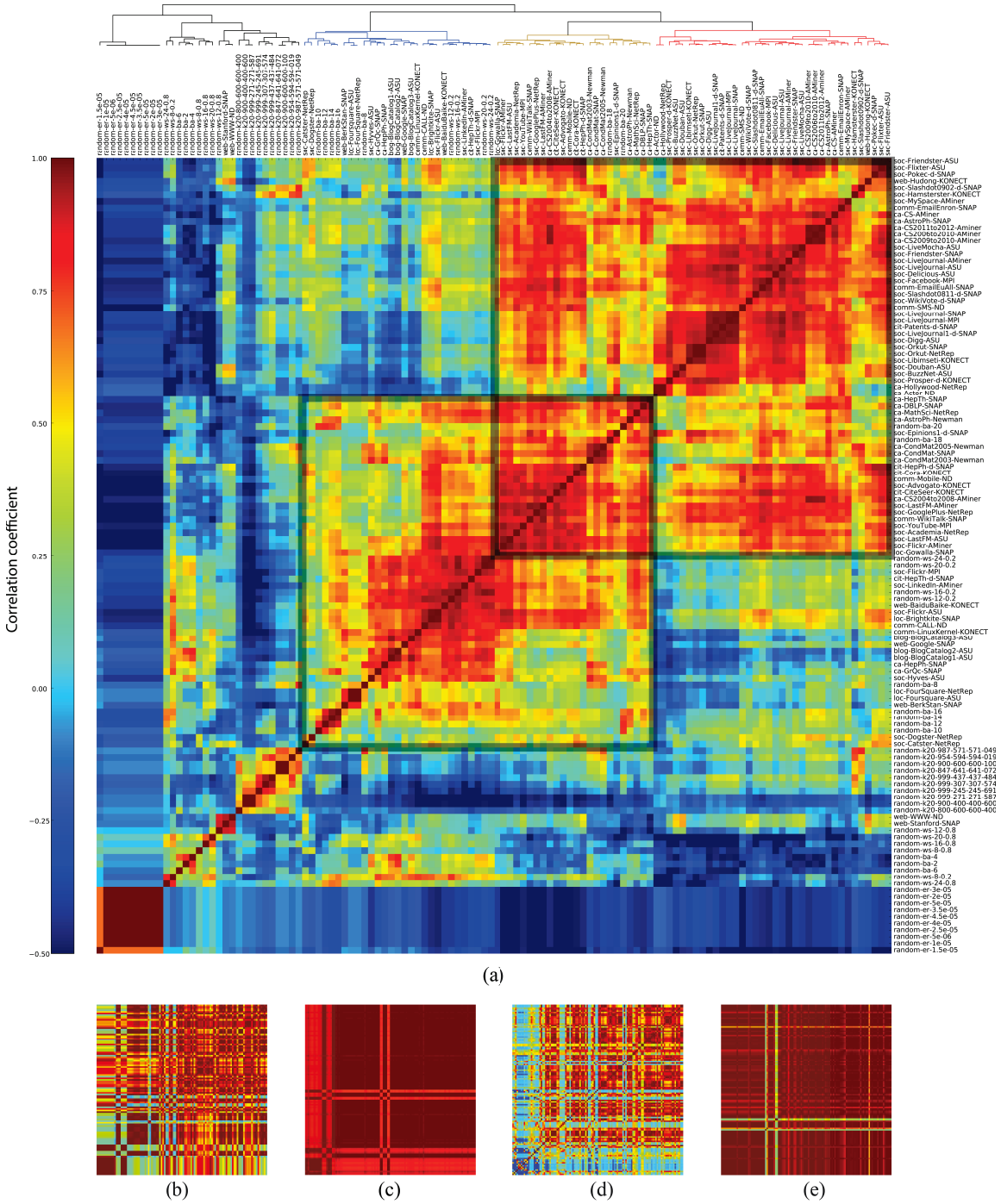



Figure 5.3. Correlation coefficient matrix of different methods for 120 networks. (a) Structural diversity signature. (b) Subgraph significance profile. (c) Sequence of percentile degrees. (d) Bag of degrees. (e) Bag of #CNs.

diversity for each pair of networks.

Figure 5.3 visualizes the similarity matrix of the structural diversity profiles between every pair of networks. The $x$- and $y$-axes represent all 120 networks studied in this work, and the spectrum color represents the correlation coefficient. Note that the arrangement of rows and columns in the presented similarity matrix is determined by the Ward variance minimization algorithm for hierarchical clustering [225], and the similarity between structural diversity profiles is measured by the Pearson correlation coefficient [148].

**Network Superfamilies.** We observe two major overlapping clusters on the top right of the matrix in Figure 5.3(a), which mostly consist of real networks. There exists several minor clusters on the lower left, which corresponds to most of the random graphs. The fact that our similarity analysis distinguishes between real and artificial networks suggests that the common neighborhood signatures capture hidden properties underlying the network structures.

According to the common neighborhood signature, a total of 36 networks—including Facebook and Friendster—are grouped together into a single cluster (colored 'red' in the dendrogram), wherein the structural diversity of common neighborhoods has similar effects on link formation in each network. Another 29 networks—including LinkedIn and BlogCatalog—are grouped into another cluster (colored 'blue'), indicating strong correlations among these networks but weak correlations between these networks and those in the red cluster. The overlap of the two clusters, if considered separately, consists of 24 additional networks (colored 'gold'). We note that the networks in the gold cluster (the overlapping part) demonstrate relatively higher similarity with networks in the red and blue clusters than the networks in the red and blue clusters demonstrate with each other. Finally, there are 31 remaining networks (2 real networks and 29 random graphs) that are not clustered into any of the three aforementioned clusters (colored 'black').

We find that the vast majority of real networks are, based on their common neighborhood signatures, clustered into three major superfamilies (colored red, blue, and gold in the dendrogram). Each superfamily consists of different types of networks; for example, the 'red' superfamily is mainly composed of social (soc-), film and academic collaboration (ca-), and email and mobile communication (comm-) networks. Another observation is that the same type of networks are grouped into different superfamilies. According to conventional wisdom, for example, the Facebook and LinkedIn networks both belong to the concept of online social networks (soc-). However, the Facebook network is indexed in the 'red' superfamily, while the LinkedIn network is indexed in the 'blue' superfamily, demonstrating that the structural diversity of common neighborhoods actually serves opposing roles in determining link existence within these two networks.

**Random Graphs.** We further study how the common neighborhood signatures qualify the nature of random graphs. Observed from Figure 5.3(a), all Erdős-Rényi (ER) graphs [59] are densely clustered into the bottom left hierarchy. According to Watts and Strogatz [228], WS random graphs with the $\beta$ parameter close to 1 tend to approach ER random graphs. This theory is captured by the structural diversity profile, as WS graphs with $\beta$=0.8 show the highest similarity with ER graphs.

What we find the most striking is that while WS and BA graphs may imitate specific superfamilies (blue or gold) real networks (e.g., BlogCatalog and LinkedIn), none of them are able to simulate an important family (red) of real-world networks (which includes networks like Facebook, Friendster, and MySpace). Even the 10 Kronecker graphs with the original parameters fitted to 10 real networks do not belong to any of the three major superfamilies, but form a standalone cluster. These observations indicate that although the random graph models studied may satisfy a series of network properties, including scale-free and small-world phenomena, the common neighborhood signature is a novel network organizational property that is

not captured by other metrices.

### 5.5.2 Network Property

Characterization of the global similarity and difference across multiple networks is conventionally focused on degree distribution [14, 62], degree sequence [61, 163], and subgraph frequency [148]. To examine the significance of the common neighborhood signature, we need to investigate not only its ability to effectively characterize networks, but the extent to which these characterizations are distinct from those provided by conventional methods. Therefore, a crucial question remains: Does the common neighborhood signature serve as a general, fundamental property of networks?

To answer this question, we analyze the common neighborhood signature at the micro and macro scales over three representative networks (as determined by the superfamilies discovered in Section 5.5.1)—namely, Friendster, BlogCatalog, and YouTube. Nevertheless, our findings generally hold for any network within a given superfamily. As there are several ways to quantify network properties at the global scale, we compare the common neighborhood signature with the following four conventional approaches: (1) The subgraph significance profile, a numerical vector of the frequencies (significance level) of different subgraphs [148]. (2) A sequence-of-percentile-degrees vector of node degrees that are ranked at particular positions (e.g., 0%, 10%, 20%, ..., 90%, 100%) of a network's degree sequence. (3) A bag-of-degrees vector of occurrence counts of node degrees in a network, which is equal to its degree distribution. (4) A bag-of-#CNs vector, in which the occurrence counts of common neighborhood size is vectorized.

At the micro scale, we can examine the visualized distributions of the aforementioned measures. The four-subgraph distributions (computed by ESCAPE [34]) are shown in Figure 5.4(b), degree distributions in Figure 5.4(a), and common neighbor-

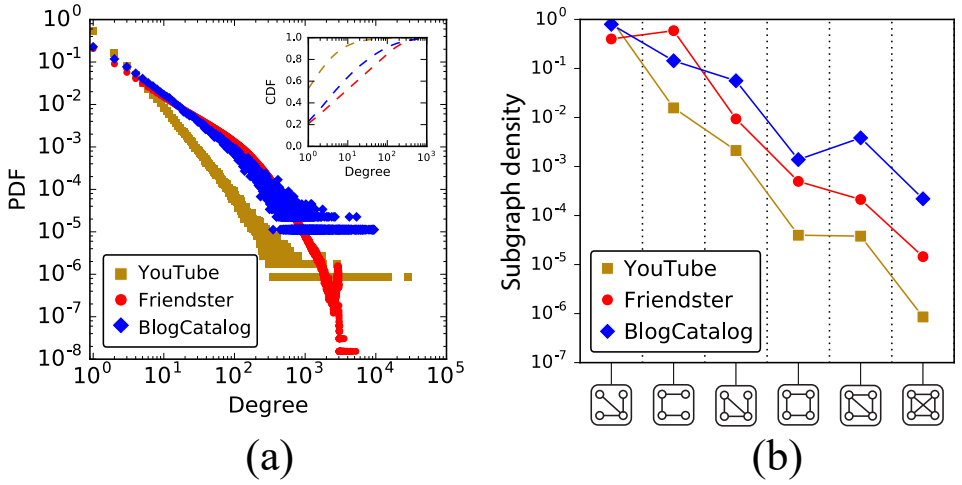



Figure 5.4. Degree distribution (a) and four-node subgraph frequency distribution (b) in three networks.

hood size distributions in Figure 5.2(b). From these figures, we can observe that, while not identical, each type of distribution reveals similar trends and shapes within the three networks. We also provide numerical results of the differences between Friendster and BlogCatalog. The correlation coefficients based on subgraph, sequence of percentile degrees, bag of degrees, and bag of #CNs are 0.579, 1.000, 0.996, and 0.957, respectively, which are significantly higher than common neighborhood signature based quantification ($-0.267$). The strong correlations between Friendster and BlogCatalog produced by all four alternative methods further highlight their inability to uncover hidden network properties.

At the macro scale, we can examine heatmaps of the correlation coefficient matrix for each method. The heatmap for the common neighborhood signature is shown in Figure 5.3(a), while the heatmap for the alternative methods are shown in Figure 5.3(b)(c)(d)(e). To compare with the common neighborhood signature, the ordering of networks in these four matrices are kept identical to that in Figure 5.3(a). The four resulting matrices fail to show clear and dense clusters, further confirming the unmatched ability of common neighborhood signatures to detect unique network

superfamilies. Note that the subgraph significance profile (Figure 5.3(b)) is able to categorize the networks into different superfamilies if the same clustering algorithm is applied to the correlation matrix. Based on these results, we argue that the common neighborhood signature is able to capture underlying mechanisms of network organization that cannot be discovered by conventional methods such as the subgraph significance profile [148], degree distribution [62], and degree sequence [163].

**Conclusion.** Our comprehensive study based on both micro- and macro-level phenomena demonstrates that the common neighborhood signature can detect intrinsic, hidden network superfamilies that are not discoverable by conventional methods. These findings suggest that the common neighborhood signature serves as a unique, fundamental property intrinsic to networks.

## 5.6 Diversity and Embeddedness in Link Existence

By leveraging the structural diversity signature, we discover three major superfamilies from the 80 real-world networks. To further understand how the structural diversity of common neighborhoods influences link existence in the three superfamilies, we focus our investigations on the following three large-scale social networks, each of which represents a particular network superfamily: **Friendster** (65,608,366 nodes and 1,806,067,135 edges) from the 'red' superfamily, **BlogCatalog** (88,784 nodes and 2,093,195 edges) from the 'blue' superfamily, and **YouTube** (1,134,890 nodes and 2,987,624 edges) from the 'gold' superfamily. (see Figure 5.3(a)). *However, our findings generally hold for any network within a given superfamily.*

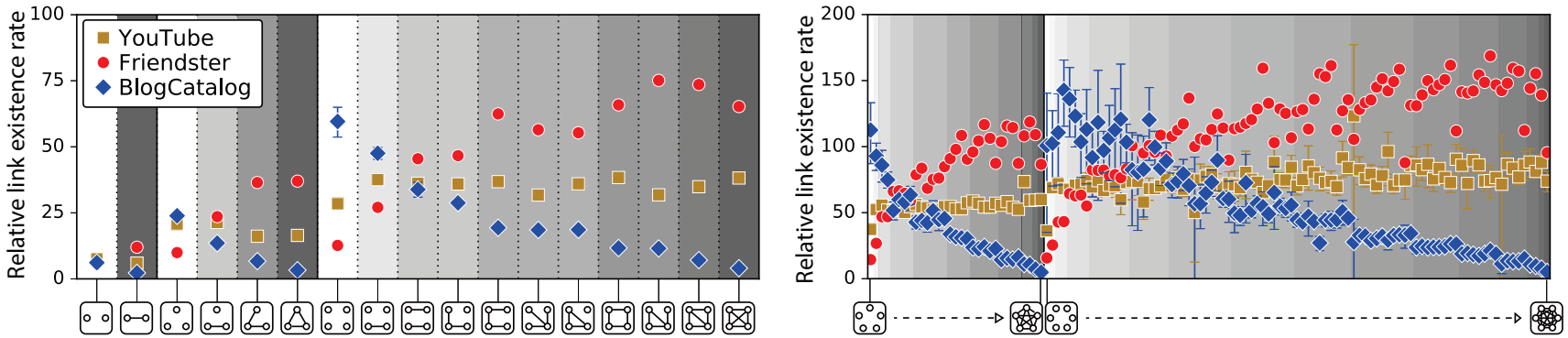



Figure 5.5. Structural diversity of common neighborhoods in link existence. The network colors—red, blue, and gold—are in accordance with the three superfamily hierarchies of the dendrogram in Figure 5.3, respectively. $x$-axis: two-node, three-node, and four-node common neighborhoods on the left side; five-node and six-node common neighborhoods on the right side. The $x$-axis is ordered according to the following keys: common neighborhood size (ascending), edge density of the common neighborhood (ascending), and component count of the common neighborhood (ascending). When all three keys are the same, the degree sequence of the common neighborhood is in descending order. Shading indicates differences in edge density. Error bars designate the 95% confidence interval.

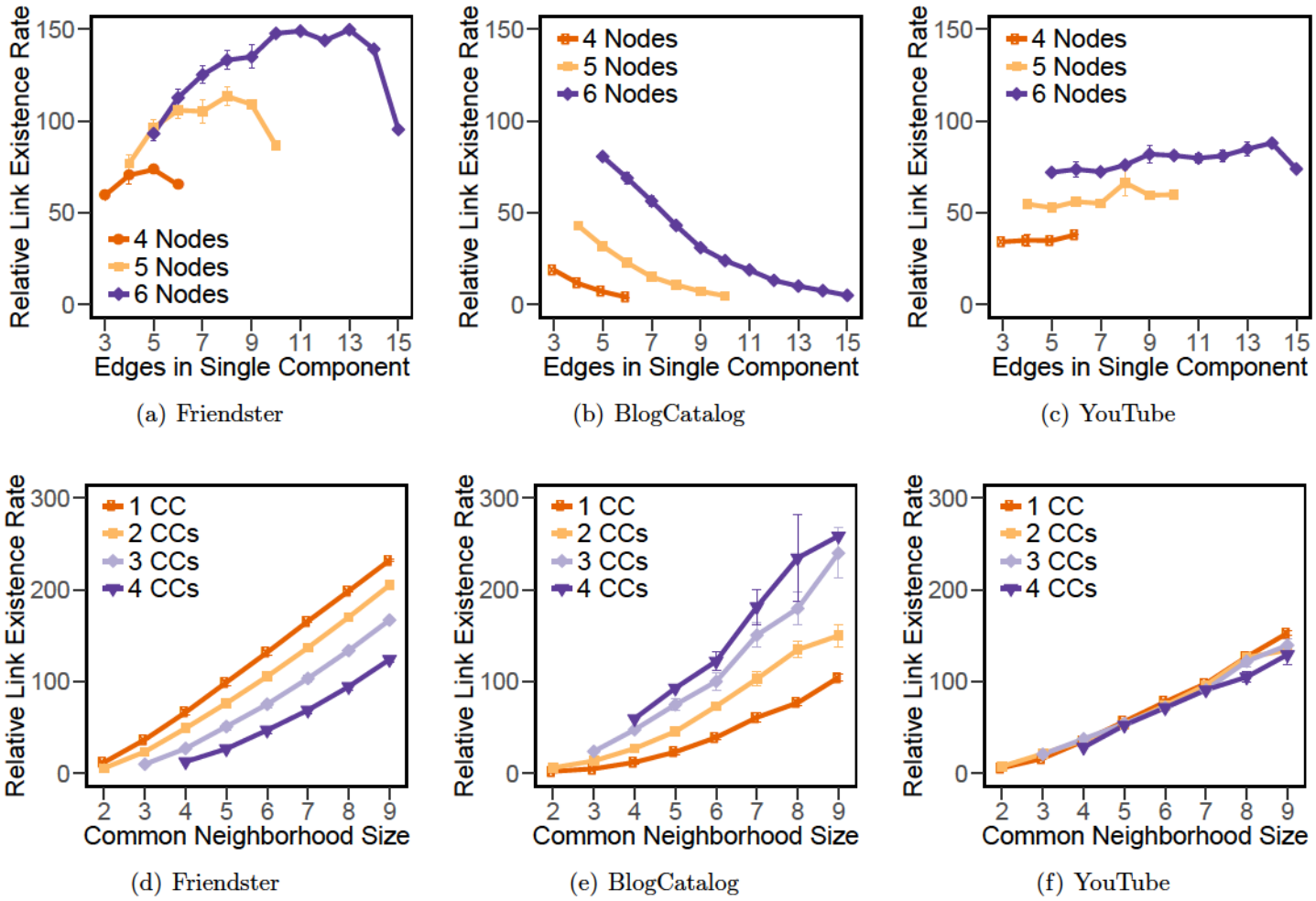
### 5.6.1 Link Existence Correlation

When we control for the size of common neighborhood between two users, how does its structure influence the probability that they form a link in the network? An illustrative example of this question is introduced as follows. Given that two users $v_i$ and $v_j$ have four common neighbors, are they more likely to connect with each other if their four common neighbors do not know each other ( ) or if their four common neighbors already know each other ( ), i.e.,

$$P(e_{ij}{=}1 \mid ) \gtrless P(e_{ij}{=}1 \mid ) \ ?$$

To address this question, we compute 546 billion, 612 million, and 1.26 billion pairs of users in the Friendster, BlogCatalog, and YouTube networks, respectively. Figure 5.5 presents the relative link existence rates in the Friendster, BlogCatalog, and YouTube networks for neighborhoods with between two and six common neighbors. It is immediately observable that the impact of structural diversity and embeddedness on link existence in the three networks is remarkably varied.

We study in detail how the structure—that is, the diversity and embeddedness—of common neighborhoods influences the link existence. Recall that embeddedness is measured by the edge count between all pairs of nodes, and diversity is measured by the number of connected components. In general, if we control the size of common neighborhood, then as the component count increases, the link existence rate decreases in Friendster (and its superfamily) but increases in BlogCatalog (and its superfamily). This finding is illustrated in Figure 5.5 and Figure 5.6(d)(e). This tells us that users on BlogCatalog are more likely to connect if their common friends are more structurally diverse, while users on Friendster are more likely to connect if their common friends are densely embedded in the same community.




(a) Friendster (b) BlogCatalog (c) YouTube

(d) Friendster (e) BlogCatalog (f) YouTube

Figure 5.6. Diversity and embeddedness vs. link existence. (a)(b)(c) Link existence rate as a function of edge count (embeddedness) with one component. (d)(e)(f) Link existence rate as a function of component count (diversity).

For common neighborhoods with the same size (fixed homophily) and component count (fixed diversity), we still observe variations in the link existence rate. We further examine the impact of edge count among common neighborhoods, as distinguished by different shadings in Figure 5.5. For a given size of common neighborhood, if we focus on common neighbor subgraphs with the same edge count (shadings), the link existence rates are relatively similar to each other. For example, when two users have four common neighbors, they have similar probabilities to connect if their common neighborhood forms the following structures:  or  (four edges),  or  or  (three edges),  or  (two edges). On the other hand, with increasing densities in common neighborhoods, as in Figure 5.5 and Figure 5.6(a)(b), the relative link existence rate increases in Friendster but decreases in BlogCatalog. Indeed, the embeddedness of the common neighborhood also has a crucial role in determining link existence in both the Friendster and BlogCatalog networks.

Recall that a common neighborhood with more components is considered more diverse. If we fix the number of common neighbors, we find that the structural diversity of common neighborhoods has a negative effect on the formation of online friendships in Friendster (and its superfamily) but a positive effect in BlogCatalog (and its superfamily). This reveals a fundamental difference between these two networks and their superfamilies both in their microscopic structures and link formation mechanisms.

Further, we quantify the impact of embeddedness and diversity on link existence. Table 5.2 reports the Pearson correlation coefficients $\rho$ between the relative link existence rate and the structural diversity and embeddedness of common neighborhoods. We can clearly see that both diversity and embeddedness are strongly correlated ($|\rho| > 0.8$) with the link existence rate in Friendster and BlogCatalog, although one is positively correlated and the other is negatively correlated.

TABLE 5.2

CORRELATION ANALYSIS FOR RELATIVE LINK EXISTENCE

| Network | #CN | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Friendster | Embeddedness | 1.0 | 0.94 | 0.89 | 0.84 | 0.81 |
| | Diversity | −1.0 | −0.99 | −0.95 | −0.88 | −0.79 |
| BlogCatalog | Embeddedness | −1.0 | −0.97 | −0.95 | −0.95 | −0.92 |
| | Diversity | 1.0 | 0.98 | 0.94 | 0.89 | 0.67 |
| YouTube | Embeddedness | −1.0 | −0.84 | 0.40 | 0.67 | 0.50 |
| | Diversity | 1.0 | 0.86 | −0.38 | −0.55 | −0.38 |

**Conclusion.** We demonstrate that the structural diversity and embeddedness of common neighborhoods are crucial factors in determining link existence across networks. Further, the contrasting influences of common neighborhood structure on the link existence correspond to networks cataloged to different superfamilies (shown in Figure 5.3). This observation reaffirms our claim that the common neighborhood signature (CNS)—the means by which we organize these superfamilies—serves as a fundamental property of networks.

### 5.6.2 Violation of Homophily

Previously, we demonstrated that in Friendster's 'red' network superfamily the structural diversity (embeddedness) of common neighborhoods is in general negatively (positively) associated with link existence, i.e., $P(e{=}1|$  $) < P(e{=}1|$  $)$, while in BlogCatalog's 'blue' superfamily it is in general positively (negatively) associated with link existence, i.e., $P(e{=}1|$  $) > P(e{=}1|$  $)$. A subsequent question one may ask is whether structural diversity and embeddedness conflict with the principle

of homophily. Specifically, for Friendster, this can be formalized as:

$$P(e{=}1|\,\boxed{\circ\circ}\,) \;>\; P(e{=}1|\,\boxed{\triangle}\,)\;?$$

In BlogCatalog, the question can be similarly formalized as:

$$P(e{=}1|\,\boxed{\times}\,) \;>\; P(e{=}1|\,\boxed{\circ\circ}\,)\;?$$

Conventional wisdom may answer "yes" to both cases, as the concept of structural homophily suggests that, all other things equal, relationships are more likely to form between individuals that share a larger common neighborhood. Surprisingly, however, we find that there is no empirical evidence to support the existence of homophily within the context of structural diversity and embeddedness.

In Figure 5.5, we can observe that the link existence rate between two BlogCatalog users with densely connected common neighbors is actually *lower* than the link existence rate between users with fewer but more loosely connected (more diverse) neighbors. For example, if two users share four common neighbors, the probability that there exists a link between them is, in more than half of the eleven configurations ($\boxed{\times}$, $\boxed{\times}$, $\boxed{\times}$, $\boxed{\square}$, $\boxed{\times}$, $\boxed{\times}$, $\boxed{\square}$), lower than the probability of a link between users that share three disconnected common neighbors ($\boxed{\circ\circ}$). In fact, $P(e{=}1|\,\boxed{\circ\circ}\,)$ is 493% higher than $P(e{=}1|\,\boxed{\times}\,)$; even $P(e{=}1|\,\boxed{\circ\circ}\,)$ is higher than $P(e{=}1|\,\boxed{\times}\,)$. By contrast, in Friendster, homophily is instead violated when a larger number of disjoint common neighbors meets with a smaller number of connected ones. For example, the link existence probabilities given $\boxed{\circ\circ}$ and $\boxed{\circ\circ}$ are lower than those given $\boxed{\triangle}$ and $\boxed{\angle}$. Similar violations can be seen to occur in various cases with different numbers of common neighborhoods.

TABLE 5.3

REGRESSION ANALYSIS FOR LINK EXISTENCE RATE

| Network | Friendster | BlogCatalog | YouTube |
|---|---|---|---|
| Intercept | −0.03845 *** | 0.00010 | −0.01855 *** |
| #CN | 0.01948 *** | 0.00252 *** | 0.00792 *** |
| Embeddedness | 0.03234 *** | −0.01580 *** | 0.00563 ** |
| Diversity | −0.01102 *** | 0.00114 *** | −0.00047 |
| Adj. $R^2$ (CNS) | 0.83330 | 0.76750 | 0.81440 |
| Adj. $R^2$ (Homophily) | 0.42300 | 0.14260 | 0.77160 |

### 5.6.3 Link Prediction

The configuration of common neighborhoods is crucial in determining link existence. Often it violates the principle of structural homophily, which demonstrates a simple and yet effective predictor for inferring link existence. Accordingly, we further explore the extent to which the diversity and embeddedness of common neighborhoods can help link inference. More formally, we ask which of the following measurements is more accurate,

$$P(e{=}1 \mid \boxtimes) \quad \text{or} \quad P(e{=}1 \mid \#\text{CN}{=}4) \ ?$$

Note that to answer this question, we focus on qualifying the effect of structural diversity and embeddedness in link inference and its potential for a new unsupervised link predictor feature, rather than targeting at the link prediction problem.

First, to demonstrate the role of CNS in link inference, we perform a regression analysis, shown in Table 5.3. We find that, together with the size of common neigh-

borhoods (#CN), the two characteristics of CNS—embeddedness and diversity—can be used as highly accurate predictors of the link existence rate ($R^2$>0.75). We also find that they serve as statistically significant ($p$<0.001) factors in the Friendster and BlogCatalog networks. Observed from the last row of Table 5.3 (Significance code: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$), when predicting for the Friendster and BlogCatalog networks, we can achieve a far better estimation by using the structural diversity of common neighborhoods than using only structural homophily (#CN), as measured by $R^2$. On Friendster, $R^2$ improves from 0.42 to 0.83 (+97%), and on BlogCatalog, $R^2$ improves from 0.14 to 0.76 (+442%).

Second, we use both structural homophily and CNS as link predictors to infer whether there exists a link between two users. For structural homophily, we use #CN as the unsupervised predictor. For structural diversity, we use the linear combination of its two characteristics—density and variety—as the predictor. For these predictions, we limit the candidate pairs of users to be inferred as those users with between two and six common neighbors. This generates more than 67 billion, 224 million, and 118 million data instances in the Friendster, BlogCatalog, and YouTube networks, respectively. We further note that the ratio between positive (existing links) and negative (non-existing links) instances is highly imbalanced in each network, resulting in difficult prediction tasks.

Table 5.4 shows the link inference performance generated by structural homophily and diversity on each of the three networks as measured by AUPR and AUROC. Figure 5.7 illustrates the corresponding precision-recall curves. In terms of AUPR, the CNS-based unsupervised predictor outperforms the homophily-based predictor by about 57% in the Friendster and BlogCatalog networks. In terms of AUROC, CNS also demonstrates greater predictive power than homophily. An application of the $t$-test to these results finds that the improvements of the diversity-based predictor over homophily-based predictor are highly statistically significant ($p \ll 0.001$).

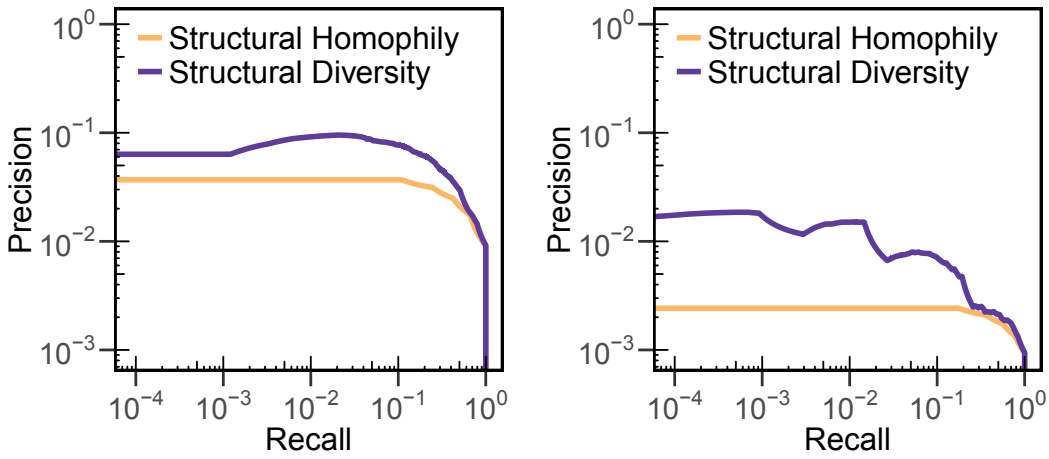
TABLE 5.4

INFERRING LINK EXISTENCE

| Metric | Method | Friendster | BlogCatalog | YouTube |
|---|---|---|---|---|
| Data | #Pairs | 67,033,108,105 | 224,786,028 | 118,635,122 |
| | %Positive | 0.9183% | 0.0943% | 0.5082% |
| AUPR | Homophily | 0.02230 | 0.00178 | 0.01524 |
| | CNS | 0.03499 | 0.00279 | 0.01532 |
| AUROC | Homophily | 0.68539 | 0.66259 | 0.69371 |
| | CNS | 0.71722 | 0.70239 | 0.68401 |

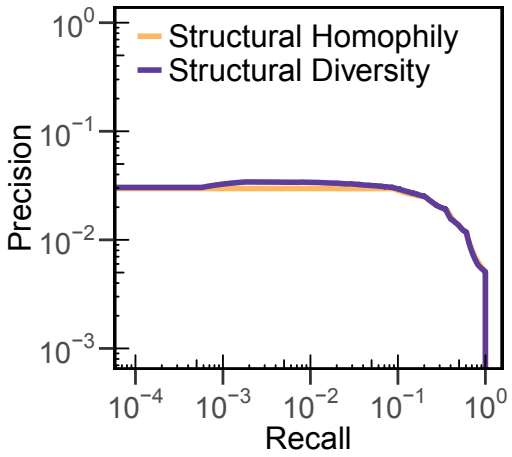
Note that the structural diversity-based predictor does not outperform the structural homophily-based predictor on the YouTube network. While the lack of improvement could be considered disappointing, this result actually further validates the findings in Figure 5.5, which shows that the impact of the diversity (or Embeddedness) of common neighborhoods on link existence can differ among networks of different superfamilies. Specifically, this result shows that the influence of CNS on networks in the 'gold' superfamily, which includes the YouTube network, is not as significant as it is for the 'blue' and 'red' superfamilies. That is, the observed difference in performance is a consequence of the underlying factors that distinguish the 'gold' superfamily from the others.

**Conclusion.** We provide empirical evidence that the structural diversity and embeddedness of common neighborhoods helps the link inference task for networks in the 'blue' and 'red' superfamilies, and we demonstrate that this performance reaffirms the existence of superfamilies. As a result, we find that the proper application of CNS has the potential to substantially improve the predictability of link existence,




(a) Friendster

(b) BlogCatalog




(c) YouTube

Figure 5.7. Precision-recall curves for inference of link existence.

with important implications for improving recommendation functions employed by social networking sites.

## 5.7 Related Work

Social theories are the empirical abstraction and interpretation of social phenomena at a societal scale. The idea of homophily, in particular, dates back thousands of years to Aristotle, who observed that people "love those who are like themselves" [10].

In its modern-day conception, the principle of homophily holds that individuals are more likely to associate and bond with similar others [115, 140]. In the context of network science, structural homophily suggests that people with more common neighbors tend to connect with each other [99, 161].

**Structural Embeddedness.** Granovetter defined embeddedness as "the extent that a dyad's mutual contacts are connected to one another" [79, 81], followed by a line of work that has demonstrated the power of embeddedness in network organization and economic development [219]. Note that sometimes this concept can be also considered "to be the number of common neighbors the two endpoints have" [57, 137]. In this work, we leverage Grannovetter's definition of embeddedness and refer to "the number of common neighbors" as structural homophily [94, 99, 161]. But even with a fixed number of edges between a dyad's common neighbors, there still exist a diverse set structures that describe common neighborhoods. How these diverse structures influence link existence remains an open question in both network and social science.

**Structural Diversity.** The concept of structural diversity was first proposed in an ego network by Ugander et al. [216], who found that the user recruitment rate in Facebook is determined by the variety of an individual's contact neighborhood, rather than the size of his or her neighborhood. Further studies show that the diversity of one's ego network also has significant influence on a user's other social decisions [64, 131]. The difference between this work and our study centers around neighborhood studies. While Ugander et al. study the variety of a single individual's contact neighborhood, we instead focus on the structural diversity of a pair of individual's common neighborhoods.

Other studies have leveraged the concept of common neighborhoods [27]. Backstrom and Kleinberg designed a new tie strength metric—dispersion—based on the connections among mutual friends, which can be used to accurately infer one's significant other (e.g., husband or wife and fiancé or fiancée) from Facebook. Their goal,

however, is to classify the type of existing social tie, while we aim to quantify the diversity of common neighborhoods of both connected and non-connected users.

**Subgraph and Network Superfamily.** Our work is also related to subgraph (motif) frequency. Milo et al. investigated the distributions of subgraph frequency across multiple types of networks, and proposed a subgraph-based significance profile for networks [148]. By leveraging this profile, they discovered several network superfamilies, whereby networks in the same superfamily display similar subgraph distributions. Recently, Ugander et al. developed a framework to investigate subgraph frequencies in real networks, which is able to characterize both the empirical as well as extremal geography of large graphs [217]. However, our work is different from subgraph mining [148, 217], graph classification [108], and the graph isomorphism problem [218]. Instead of these topics, our focus is on uncovering the principles that drive the formation of local network structure and exploring the significance of structural diversity in driving link organization and network superfamily detection.

Finally, the structural diversity of common neighborhoods also offers substantial potential for applications to other important network mining tasks, including social recommendation [102], tie strength modeling [11, 80, 232], structural hole [129], community detection [122], and network evolution [121]. Structural diversity also has connections with heterogeneous network analysis [196], wherein diversity and embeddedness can be measured among different types of nodes and links.

## 5.8 Conclusion

Through this chapter, we study how the different common neighborhood structures can influence network configurations, and we examine the implications of these observations for how we organize networks. Through a comprehensive study of 120 big real-world and random networks, we conclude that, controlling for the number of common neighbors, the structure of common neighborhoods—particularly their

diversity and embeddedness—exhibits substantial influence on link existence rates. Surprisingly, although the principle of homophily has been acknowledged to hold over a wide variety of networks, we find that common neighborhood structure demonstrates properties that conflict with that of homophily.

We further define the common neighborhood signature (CNS), which serves as a fundamental property of a network similar to degree distribution, degree sequence, and subgraph distribution, but which classify the 100+ networks into three unique superfamilies not discoverable by conventional properties. Strikingly, we find that LinkedIn and Facebook belong to different network superfamilies. For example, when the size of common neighborhood is fixed to 3, an increase of its embeddedness negatively impacts the formation of professional relationship in LinkedIn and its superfamily—that is, $P(e{=}1|$  $) < P(e{=}1|$  $)$—while it actually facilitates the formation of online friendship in Facebook and its superfamily—that is, $P(e{=}1|$  $) < P(e{=}1|$  $)$, signifying important implications for "PYMK" in both services.

Furthermore, a study of representative random graph models (ER, WS, BA, and Kronecker) shows that none of the models is able to simulate a particular superfamily of real-world networks (which involves Facebook, Friendster and so on). This not only demonstrates the power of the CNS as a new, fundamental property of networks, but also provides new opportunities and suggestions for building random graph models.

Our next step will be to extend our examination of common neighborhood structure beyond homogeneous and static networks, further including heterogeneous networks and dynamic, inter-genre, attributed networks. In addition, we would like to examine the interrelations between the two characteristics of common neighborhoods—diversity and embeddedness. Finally, we intend to incorporate the CNS into machine learning frameworks to improve social recommendation performance.

# CHAPTER 6

# TOPIC DIVERSITY AND AUTHORITY

## 6.1 Overview

A widely used measure of scientific impact is citations. However, due to their heavy-tailed distribution, citations are fundamentally difficult to predict. Instead, to characterize scientific impact, in this chapter we address two analogous questions asked by many scientific researchers: "How will my $h$-index evolve over time, and which of my previously or newly published papers will contribute to it?" To answer these questions, we perform two related tasks. First, we develop a model to predict authors' future $h$-indices based on their current scientific impact. Second, we examine the factors that drive papers—either previously or newly published—to increase their authors' predicted future $h$-indices. By leveraging relevant factors, we can predict an author's $h$-index in five years with an $R^2$ value of 0.92 and whether a previously (newly) published paper will contribute to this future $h$-index with an $F_1$ score of 0.99 (0.77). We find that topical authority and publication venue are crucial to these effective predictions, while topic popularity and diversity are surprisingly inconsequential. Further, we develop an online tool that allows users to generate informed $h$-index predictions. This chapter demonstrates the predictability of scientific impact, and can help scholars to effectively leverage their position of "standing on the shoulders of giants."

This chapter is largely extracted from previous publications [45, 50]. It is a joint work with Reid A. Johnson and Nitesh V. Chawla.

## 6.2 Introduction

Scientific impact plays a pivotal role in the evaluation of the output of scholars, departments, and institutions. Scientific researchers generate scientific impact through novel discoveries and developments, which are traditionally disseminated to a wider community via publications. The impact of each of these findings and corresponding publications—both to a field of research and, by extension, to the reputation of the author—can be affected by a variety of factors, which may be directly or indirectly related to the findings themselves. Due to the confluence of such factors, a researcher's body of work is likely to be composed of findings and publications of varying impact. Consequently, it can be challenging to predict a researcher's future impact and the influence of any particular publication on this impact, regardless of how impact is measured.

Often a researcher's total number of citations is used as a measure of impact, while a researcher's total number of publications is used as a measure of productivity. However, while these simple measures are intuitive and can be useful, they also have significant limitations. For example, a solitary well-cited, impactful paper can skew the total number of citations, potentially distorting its use as a measure of overall impact. Similarly, the total number of publications can be increased by a large number of poorly cited papers, which may not be indicative of the actual productivity involved. Moreover, as citations demonstrate a heavy-tailed distribution, with the vast majority of publications receiving few citations, these simple measures are exceedingly difficult to estimate using traditional regression analysis [32, 171]. Thus, determining how many citations a given researcher or a given paper will receive is often ineffective in practice.

In light of these difficulties and limitations, we instead address two analogous questions asked by many academic researchers: "***How will my h-index evolve over time, and which of my previously and newly published papers will***

***contribute to my future h-index?***"

These questions are based on the $h$-index. As described by J. E. Hirsch, by whom the index was proposed: "A scientist has index $h$ if $h$ of his or her papers have at least $h$ citations each, and the other papers have no more than $h$ citations each" [90]. The $h$-index is thus a function of the number of publications (quantity) and the number of citations per publication (quality). As a result of its simplicity and predictive value, the $h$-index has become a *de facto* standard for measuring scientific impact.

**Present Work.** To tackle the questions of how one's $h$-index will evolve over time and which publications will contribute to it, we formulate two scientific impact prediction problems, as shown in Figure 6.1. Our first task is to predict authors' future $h$-indices based on their current scientific impact, which has been explored with data on a small sample of neuroscientists [2]. We then determine whether a given paper will influence a particular author's predicted future $h$-index, which we formalize as our primary scientific impact prediction problem. Accordingly, our second (primary) prediction problem is to determine whether a given previously or newly published paper will, after a predefined timeframe, increase the *future* $h$-index of its primary author (i.e., the paper's first author or the author with the highest $h$-index). The predicted future $h$-indices generated by the first task are used as the future $h$-indices in our primary task. Thus, in our primary task, an author's future $h$-index represents the author's expected $h$-index after the predefined period of time, with the purpose of accounting for the change in the author's $h$-index over the prediction timeframe.

**Contributions.** This chapter discerns the impact of a given publication on the primary author's $h$-index. First, we investigate the factors that influence the development of an author's scientific impact, for which we generate a model to infer an author's future $h$-index. Second, by using the future $h$-index predicted by this model

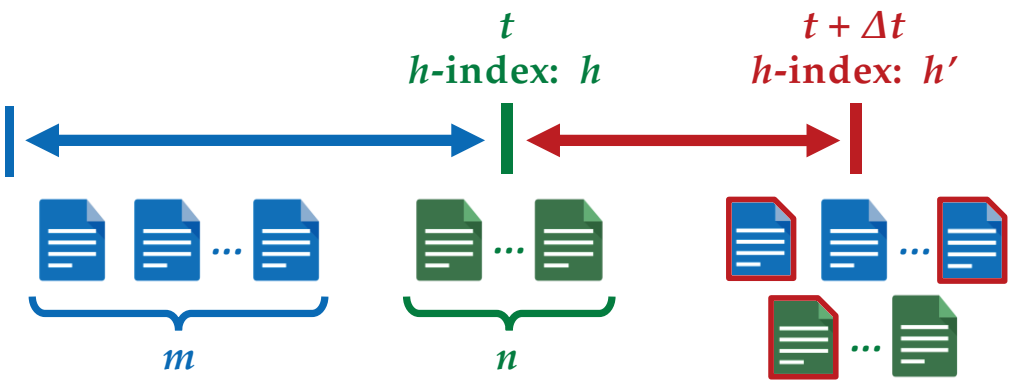



Figure 6.1. Illustrative example of scientific impact prediction. Before time $t$, a scholar published $m$ papers and had an $h$-index of $h$. Our prediction problems are targeted at answering two questions: 1) What is the scholar's future $h$-index, $h'$, at time $t+\Delta t$? 2) Which of his/her papers, both (a) those $m$ papers previously published before $t$ and (b) those $n$ new papers published at $t$, will contribute to $h'$?

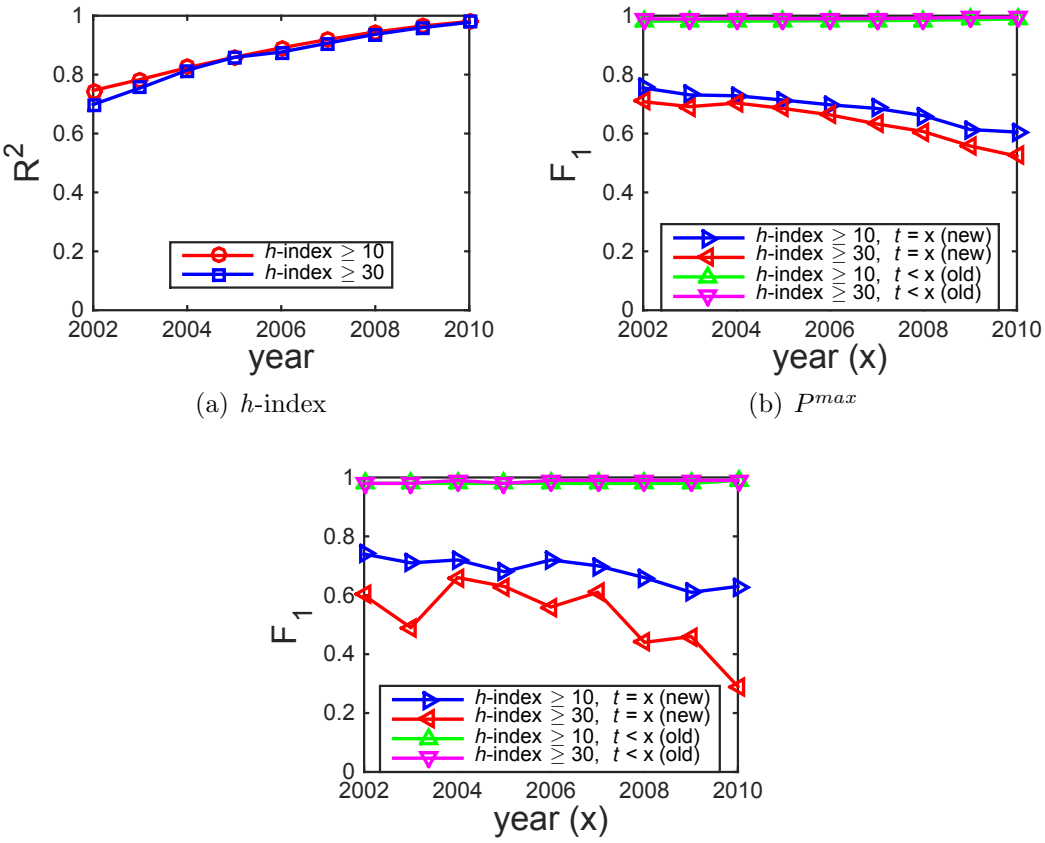
as the target variable for predicting whether a paper will increase its primary author's $h$-index, we account for the dynamic change in the primary author's $h$-index over the course of prediction timeframe. In other words, in this work we aim to predict not only on the newly published papers [45], but also on the previously published ones. We also re-define the primary author of a publication as both the first author and the author with highest $h$-index among the author list. To further add to the utility of this work, we have also developed and deployed an online tool that allows users to generate $h$-index predictions based on our findings.

**Challenges.** Factors such as the researcher's current influence, the publication topic, and the publication venue may, among many other factors, play a role in determining the degree to which a publication contributes to the researcher's future impact. A resulting challenge is the interplay of such factors, which can confound attempts to generate effective predictions. Considerations such as the variability of the $h$-index according to the "academic age" of a researcher, the widely differing citation conventions among different fields, and the co-authorship of researchers with

differing $h$-indices can make it difficult to isolate the degree to which a given paper will contribute to the measured impact of its authors. Further, effectively predicting whether a publication will contribute to its authors' measured *future* impact must account for the change in impact over the prediction timeframe, which may follow a trajectory and rate particular to each author. Our work focuses on addressing and overcoming each of these issues to generate novel, effective scientific impact predictions, as well as investigating precisely what role a variety of factors play in these predictions.

**Results.** We demonstrate a high level of predictability for scientific impact as measured by our two problems. Accordingly, we find strong performance for our first task of predicting an author's future $h$-index. Our results demonstrate that we can predict an author's $h$-index in five years with an $R^2$ value of 0.9197, as shown in Figure 6.2(a). This performance generally increases as the prediction timeframe is shortened, with a prediction of ten years achieving an $R^2$ of 0.7461. We also find strong performance for our primary task of predicting whether a publication will contribute to its primary author's future $h$-index. Our results demonstrate that we can predict whether in five years a previously (newly) published paper will contribute to the future $h$-index of the author with highest $h$-index with an $F_1$ score of 0.99 (0.77), as shown in Figure 6.2(b), an improvement of +130% (+160%) over random guessing. From Figure 6.2(c), we can observe that similar, strong performance is achieved when considering the first author of a publication as its primary author. Predictive performance for newly published papers generally increases as the prediction timeframe is expanded. However, predictive performance for previously published papers achieves consistently high $F_1$ scores, suggesting their general predictability. Our results also indicate that authors with low $h$-indices are easier to predict for than those with high ones (see Figures 6.2(b) and 6.2(c), blue vs. red lines).

We also assess the influence of various factors on our predictive results. For




(a) $h$-index (b) $P^{max}$

(c) $P^{first}$

Figure 6.2. Predictability of scientific impact. $x$-axis: year of data used to predict to 2012. $y$-axis: performance. (a) Performance for predicting an author's $h$-index as a regression task ($R^2$ value). (b) Performance for predicting whether a given paper will increase the $h$-index of its primary author (as defined by the author with highest $h$-index among its author list) as a classification task ($F_1$ score). (c) Performance for predicting whether a paper will increase the first author's $h$-index.

our first problem, predicting an author's future $h$-index, we find that the author's current $h$-index is the most important, followed by the number of publications and co-authors. For our primary problem, predicting whether a paper will contribute to its primary author's $h$-index, we find that topical authority is the most telling factor for newly published papers, while the existing citation information is the most telling for previously published ones, followed by the authors' influence and the publication

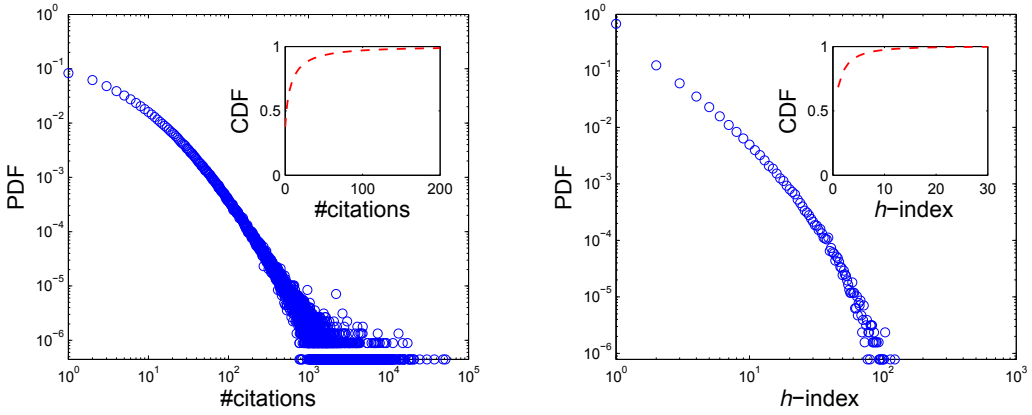
venue. We also find that the venue in which the paper is published and the author's collaborations are moderately significant factors over longer prediction periods, but become inconsequential for shorter ones. Finally, we are surprised to find that the diversity and popularity of the publication topic have no discernible correlation to the prediction target for both previously and newly published papers. Overall, our findings unveil the predictability of scientific impact and provide researchers with concrete suggestions for expanding their scientific influence and, ultimately, for more effectively "standing on the shoulders of giants."

A caveat of this work is that by targeting the $h$-index, our findings may result in unintended side effects by a principle referred to as Goodhart's Law, which essentially warns that "when a measure becomes a target, it ceases to be a good measure" [195]. Yet, we strongly believe that by deepening the understanding of scientific impact measures, the findings presented in this work can actually help to strengthen the foundations upon which these measures are based, ultimately facilitating their improved use. *In no way should our research be construed as advocating the use of the h-index or any other measure as a deciding factor in the determination of one's research pursuits.*

## 6.3 AMiner Academic Data

In this paper, we use the real-world academic dataset from ArnetMiner [203], which is the world-leading free online service for academic social network analysis and mining. The dataset contains 1,712,433 authors with 2,092,356 papers from computer science venues held until 2012. Each paper includes information on the title, abstract, authorship, references, and publication venue and year. The dataset also captures 4,258,615 collaboration (co-authorship) relationships and 8,024,869 citation relationships.

We briefly explore and report the data characteristics of the author-paper-citation




(a) Distribution of citations counts

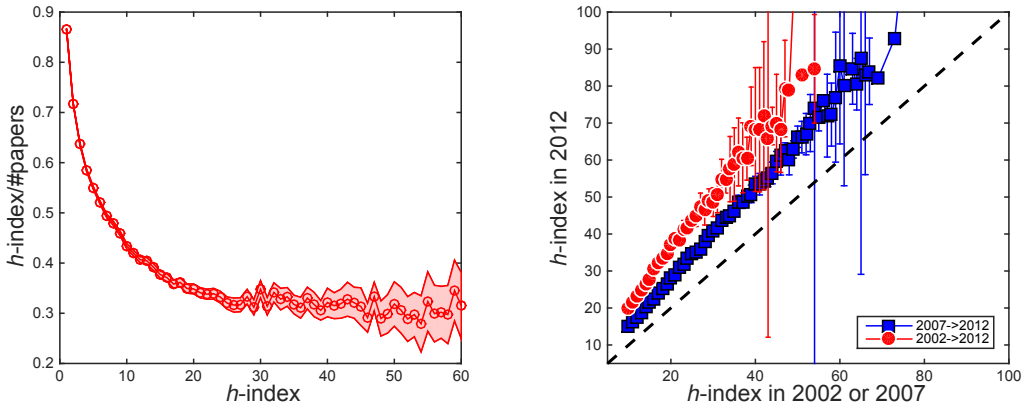
(b) $h$-index distribution

Figure 6.3. Distributions of the citation counts of papers and the $h$-indices of authors. In this dataset, 7.41% (154,985) of the papers obtain more than 50 citations and 0.0093% (159) of the researchers have $h$-indices greater than 60.

data used in this work. Figure 6.3 shows the distributions of the number of citations for each paper and the $h$-index of each author. In our dataset, both metrics follow heavy-tailed distributions (i.e., distributions with a "tail" that is "heavier" than that of an exponential). Moreover, only 7.41% (154,985) of the papers have more than 50 citations, while 0.0093% (159) of the researchers have an $h$-index over 60.

## 6.4 Problem Definition

Traditionally, the task of scientific impact prediction is formulated as a regression problem for predicting citation counts [234]. However the intrinsically heavy-tailed distribution of citation counts, demonstrated in Figure 6.3(a), make such predictions necessarily skewed [32, 45]. This problem motivates a search for alternate approaches that are more resilient to a skew in citation counts. Inspired by the work of [32], which considers the problem of Facebook cascade growth prediction, we formulate the following task: Given a paper at timestamp $t$, we predict whether that paper will increase its authors' $h$-indices by the future timestamp $t + \Delta t$.




(a) $h$-index vs. $h$-index/#papers (b) $h$-index in 2012 vs. 2002&2007

Figure 6.4. $h$-index trends. (a) The ratio between one's $h$-index ($\geq 20$) and her/his number of papers stabilizes at 0.3. (b) The correspondence between one's $h$-index in 2002 (red line) and 2007 (blue line) and his/her predicted $h$-index in 2012.

Realistically, however, the authors' $h$-indices are not static; they may increase during the duration $\Delta t$. Figure 6.4(b) shows the comparisons between scholars' $h$-indices in 2002 or 2007 and their corresponding future $h$-indices in 2012. In this sense, to solve the scientific impact prediction task above, we need to first infer the future $h$-indices of the paper's authors. Thus we formalize two prediction problems, namely future $h$-index prediction and scientific impact prediction.

**Problem 3 (Future $h$-index Prediction)** *Given the publication corpus $C$ before timestamp $t$ and each author's h-index at $t$, the task is to predict the authors' future h-indices at timestamp $t + \Delta t$.*

**Definition 5 (Primary Author)** *Given a paper $d \in C$, the primary author of $d$ is defined in two ways: given paper $d$'s author list, take either the author with the highest h-index or the first author on the list.*

**Problem 4 (Scientific Impact Prediction)** *Given the publication corpus $C$ before timestamp $t$, each paper $d \in C$ published by (at or before) $t$, and the primary*

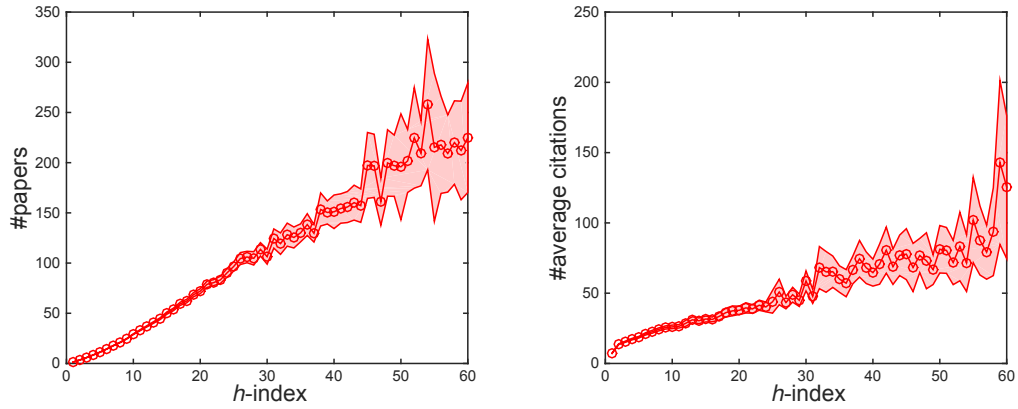
*author's predicted future h-index, the problem is to predict whether d's number of citations will reach the primary author's future h-index after a given time period $\Delta t$.*

The major novelty of this approach lies in the formulation of the second problem, i.e., scientific impact prediction, while the first problem serves to facilitate it. As formulated, the second problem is composed of two tasks. The first task is to predict for papers published before the current timestamp $t$. For these papers, we have citation counts that have accumulated until $t$. The second task is to predict for those papers published at $t$ without prior information about their citations. Importantly, the problem addresses the above-noted issues with traditional citation count prediction by using a local threshold—the primary author's $h$-index—for each paper's future citation count. Figure 6.4(a) shows that the ratio between one's $h$-index ($\geq 20$) and his or her number of papers stabilizes at about 30%, allowing us to circumvent the inherent skew of citation counts.

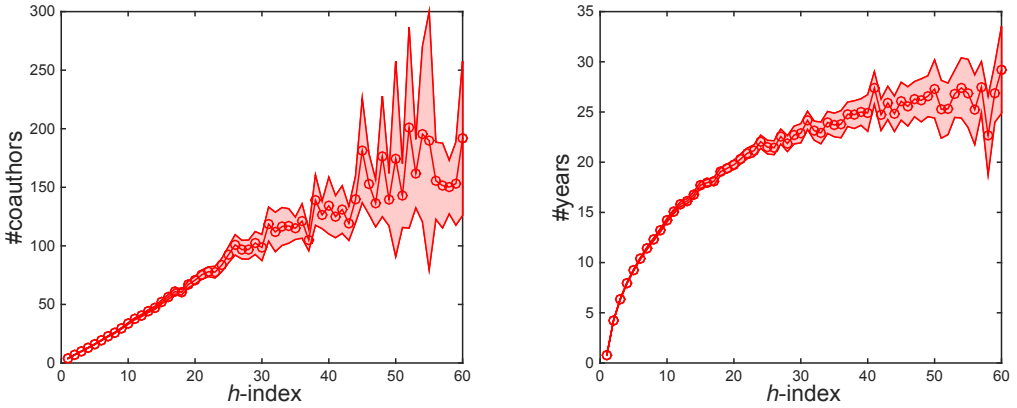
Our proposed problem of scientific impact prediction is fundamentally different from the traditional problem of predicting citation counts [234]. Whereas citation count prediction typically employs regression to predict scientific impact, our problem is to instead predict each paper's future impact conditioned on its authors. Though inspired by it, our problem is also entirely different from the cascade growth prediction problem [32], which requires the observation of the first $k$ reshares (here, citations) to predict future reshare counts. The chief advantage of our formulation is its general applicability to a variety of real-world tasks, including author $h$-index and popularity prediction [181], expert finding and search [239], and credit allocation [107, 182].

## 6.5 Scientific Impact Factors

To quantify scientific impact, it is natural to use the number of citations obtained by each paper and its authors. Recall that given a paper $d$, our objective is to predict whether the number of citations $c_d$ it obtains within a given time period $\Delta t$ will be




(a) $h$-index vs. #papers (b) $h$-index vs. #average citations




(c) $h$-index vs. #co-authors (d) $h$-index vs. #years

Figure 6.5. $h$-index factor correlations. (a) (c) The numbers of papers and co-authors are highly correlated with a scholar's $h$-index. (b) The average number of citations for each author is larger than her/his $h$-index. (d) The rate at which the $h$-index increases itself increases as the length of time spent in academia becomes longer (*i.e., the rich get richer*). Shaded area indicates error bars observed at a 95% confidence interval.

larger than its primary author's future $h$-index. In other words, we aim to model the co-evolution of the primary author's $h$-index and paper $d$'s citation count over the period $\Delta t$.

TABLE 6.1

H-INDEX FACTOR DEFINITIONS

| Factor | Description | $cc_{2002}$ | $cc_{2007}$ |
|---|---|---|---|
| *h-index* | Current $h$-index | 0.7838 | 0.9335 |
| *num-papers* | #papers published | 0.6518 | 0.7375 |
| *num-citations* | Average #citations per paper | 0.1486 | 0.2289 |
| *num-co* | #unique co-authors | 0.5784 | 0.5992 |
| *num-years* | #years since first paper | -0.0855 | 0.1089 |

### 6.5.1 Factors That Drive One's $h$-index to Increase

We first examine the factors that potentially affect the development of scientific scholars' $h$-indices. Acuna et al. [2] and Redner et al. [173] have examined the factors that are indicative of the future $h$-indices of small groups of physicists and neuroscientists, respectively. As our work focuses on the computer science domain, Table 6.1 provides brief descriptions for five simple factors that we find to have effects on the evolution of computer scientists' $h$-indices, as well as the correlation coefficients between these factors in 2002 ($\Delta t$=10 years) / 2007 ($\Delta t$=5 years) and the scholars' future $h$-indices in 2012. $cc_{2002}$ and $cc_{2007}$ represent the respective correlation coefficients.

The correlation coefficients provide several observations. First, we can observe that researchers' future $h$-indices are highly correlated with their current $h$-indices, followed by their number of publications and co-authors. Second, we notice a potentially counterintuitive phenomenon, wherein the number of citations and years publishing work have surprisingly limited correlations with future $h$-indices vis-à-vis other factors. Finally, within a shorter timeframe ($cc_{2002}$ vs. $cc_{2007}$), historical and

future $h$-indices exhibit high correlations.

Figure 6.5 presents the basic characteristics of scientific impact in terms of $h$-index, including counts for an author's number of papers, citations, co-authors, and years conducting research. Positive linear relationships are clearly observed between the $h$-index and the number of papers and co-authors in Figures 6.5(a) and 6.5(c), respectively. Also, Figure 6.5(b) shows that the average number of citations for each author is larger than his or her $h$-index. Finally, in Figure 6.5(d), we examine the interplay between authors' $h$-indices and the length of time they spend in academia (the date difference between one's first and last publications). We observe that the increase of $h$-index is relatively slow upon initially entering academia. As one's $h$-index increases, the accumulations of influence, resources, connections, and publications further drive one's $h$-index upward, and scientific impact expands at an increasingly rapid rate. In other words, the aphorism that "the rich get richer" is readily observed in academia, whereby the influence of individuals who have already accumulated a great deal of influence increases at a disproportionally quick rate. All characteristics are observed at a 95% confidence interval.

### 6.5.2 Factors That Drive Papers to Increase $h$-index

We further investigate the factors that drive a paper's citation count to exceed its primary author's $h$-index, including the paper's author(s), content, publication venue, and references, as well as social and temporal effects related to its author(s). Table 6.2 lists the six diverse groups of factors investigated in this work, and Table 6.3 reports the correlation coefficients between the factors of papers published in 2002 ($\Delta t = 10$) / 2007 ($\Delta t = 5$) and whether their citation counts are greater than or equal to the primary authors' $h$-indices in 2012. Figure 6.6 shows the response curve of the most important factor for each group (as evaluated by correlation coefficients in Table 6.3) when considering the max-$h$-index author as the primary author.

TABLE 6.2

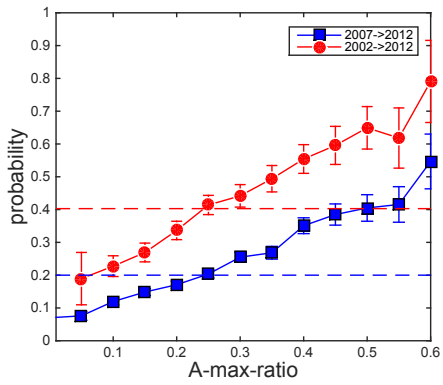
FACTOR DEFINITIONS

| Category | Factor | Description |
|---|---|---|
| Author | *A-first-max* | The first author's $h$-index. |
| | *A-ave-max* | The average $h$-index of all authors. |
| | *A-sum-max* | The sum of $h$-indices of all authors. |
| | *A-first-ratio* | The ratio between max-$h$-index and #papers attributed to the first author. |
| | *A-max-ratio* | The ratio between max-$h$-index and #papers attributed to the primary author. |
| | *A-num-authors* | The number of authors of the given paper. |
| Content | *C-popularity* | The #average-citations over different topics (see Eq. 6.1). |
| | *C-novelty* | The topic novelty of this paper (see Eq. 6.2). |
| | *C-diversity* | The topic diversity of this paper (see Eq. 6.3). |
| | *C-authority-first* | The consistence between the first author's authority and this paper (see Eq. 6.4). |
| | *C-authority-max* | The consistence between the primary author's authority and this paper. |
| | *C-authority-ave* | The average consistence between each author's authority and this paper. |

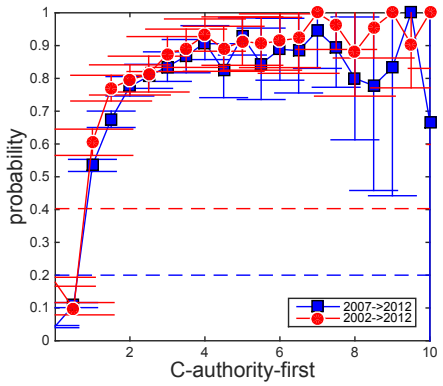
TABLE 6.2

*Continued*

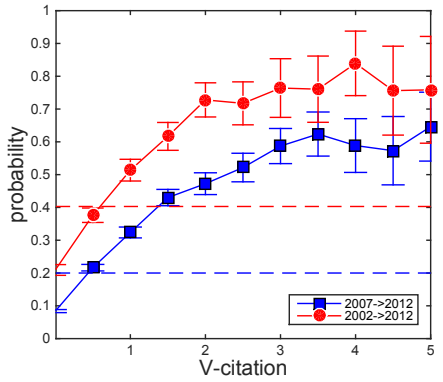
| Category | Factor | Description |
|---|---|---|
| Venue | *V-h-index* | The venue's $h$-index. |
| | *V-citation* | The #average-citations of papers published in this venue. |
| Social | *S-degree* | The number of co-authors of the paper's authors. |
| | *S-pagerank* | The PageRank values of the paper's authors in the weighted collaboration network. |
| | *S-h-coauthor* | The average $h$-index of co-authors of the paper's authors. |
| | *S-h-weight* | The weighted average $h$-index of co-authors of the paper's authors. |
| Reference | *R-h-index* | The references' $h$-index. |
| | *R-citation* | The #average-citations. |
| Temporal | *T-ave-h* | The average $\Delta h$-indices of the authors between now and three years ago. |
| | *T-max-h* | The maximum $\Delta h$-index between now and three years ago. |
| | *T-h-first* | The $\Delta h$-index of the first author between now and three years ago. |
| | *T-h-max* | The $\Delta h$-index of the max-$h$-index author between now and three years ago. |



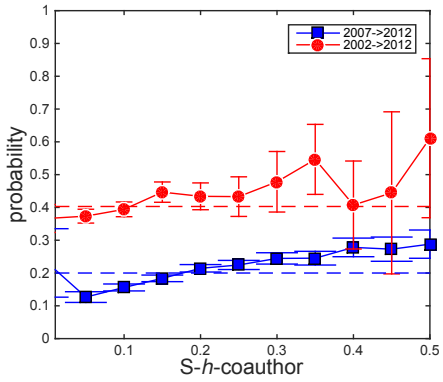

(a) Author factors



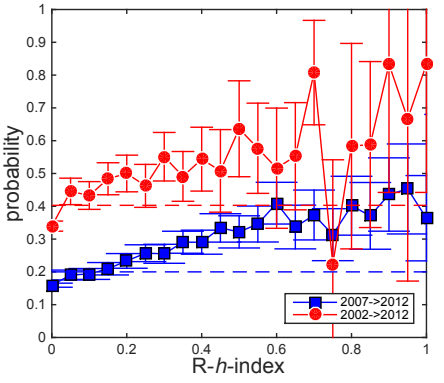

(b) Content factors



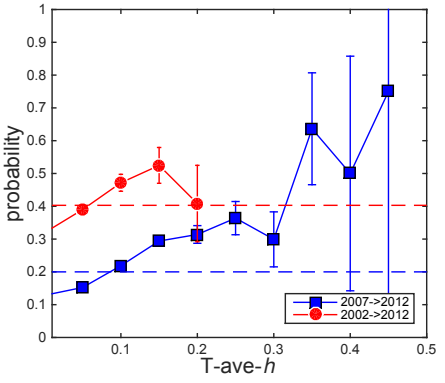

(c) Venue factors




(d) Social factors




(e) Reference factors




(f) Temporal factors

Figure 6.6. Factor response curves with $\Delta t = 5$ or 10 for $P_{new}^{max}$. $x$-axis: factor value; $y$-axis: probability that a paper published at time $t$ will increase its primary author's $h$-index by 2012. All response probabilities are observed at a 95% confidence interval.

TABLE 6.3

FACTOR CORRELATIONS

| Category | Factor | $P_{new}^{max}$ | | $P_{new}^{first}$ | |
|---|---|---|---|---|---|
| | | $cc_{2002}$ | $cc_{2007}$ | $cc_{2002}$ | $cc_{2007}$ |
| Author | *A-first-max* | 0.0309 | 0.0728 | 0.1102 | 0.1998 |
| | *A-ave-max* | 0.0435 | 0.0999 | 0.0670 | 0.0264 |
| | *A-sum-max* | 0.1589 | 0.1585 | 0.1801 | 0.1915 |
| | *A-first-ratio* | 0.0161 | -0.0365 | 0.2904 | 0.3232 |
| | *A-max-ratio* | 0.2866 | 0.2423 | 0.2601 | 0.2285 |
| | *A-num-authors* | 0.0878 | 0.0617 | 0.1359 | 0.0668 |
| Content | *C-popularity* | 0.2085 | 0.0741 | 0.2590 | 0.0628 |
| | *C-novelty* | 0.1192 | 0.0807 | 0.1262 | 0.0763 |
| | *C-diversity* | 0.1852 | 0.0712 | 0.2498 | 0.0716 |
| | *C-authority-first* | 0.3537 | 0.4346 | 0.3408 | 0.3490 |
| | *C-authority-max* | 0.3265 | 0.3874 | 0.3420 | 0.3667 |
| | *C-authority-ave* | 0.3611 | 0.4359 | 0.3623 | 0.3865 |
| Venue | *V-h-index* | 0.2557 | 0.2940 | 0.2400 | 0.2351 |
| | *V-citation* | 0.3357 | 0.3506 | 0.3058 | 0.3194 |
| Social | *S-degree* | 0.0314 | -0.0393 | 0.0340 | 0.0454 |
| | *S-pagerank* | -0.0341 | -0.0782 | 0.0500 | 0.1317 |
| | *S-h-coauthor* | 0.0750 | 0.0976 | 0.0148 | 0.0206 |
| | *S-h-weight* | 0.0639 | 0.0861 | 0.0006 | 0.0166 |
| Reference | *R-h-index* | 0.1405 | 0.1562 | 0.1204 | 0.1103 |
| | *R-citation* | 0.0858 | 0.0420 | 0.0635 | 0.0150 |

TABLE 6.3

*Continued*

| Category | Factor | $P_{new}^{max}$ | | $P_{new}^{first}$ | |
|---|---|---|---|---|---|
| | | $cc_{2002}$ | $cc_{2007}$ | $cc_{2002}$ | $cc_{2007}$ |
| Temporal | *T-ave-h* | 0.2528 | 0.2616 | 0.1740 | 0.1819 |
| | *T-max-h* | 0.2539 | 0.2027 | 0.2426 | 0.2032 |
| | *T-h-first* | 0.2109 | 0.2188 | 0.1737 | 0.0907 |
| | *T-h-max* | 0.2117 | 0.1504 | 0.2012 | 0.1603 |

**Author Factors.** The prediction task for each paper naturally depends on the authors themselves, including both the primary author and his or her co-authors. Prior work has been devoted to examining the interplay between scientific impact (number of citations) and the average values of authors' attributes [28, 234]. Given our problem formulation, in addition to these factors, for each paper we investigate the attributes of the primary author (e.g., the ratio of the author's previous papers that contribute to his/her $h$-index). Additionally, as the first author of a publication usually leads the collaboration and may have considerable influence on its scientific impact, we consider the probability that the number of citations obtained by each of the first author's previous publications is greater than the primary author's $h$-index. Of course, as a paper is the sum of all authors' contributions, the combined impact of all co-authors may influence a paper's quality and popularity. Thus the sum of all authors' $h$-indices is used to simulate their overall impact. Due to self-citation behavior, the author's productivity (i.e., the number of her/his previous publications) also has a positive effect on the paper's future citation counts [16].

**Content Factors.** Aside from the attributes of its authors, another intuitive factor affecting a paper's success is its content. Topic modeling is a widely used method for extracting and mining the content of literature and can be used to extract "topics" that occur in a collection of documents. One of the most popular topic modeling methods is known as Latent Dirichlet Allocation (LDA), a generative probabilistic approach that views each document as a mixture of various topics [19]. Similar to previous work on modeling citation counts [234], we run a 100-topic LDA model on the title and abstract of the corpus $C$ published before time $t$ and the target papers published at time $t$, which returns the probability distribution $p(z|d)$ over topics $z \in Z$ assigned for each paper $d$. We denote a target paper $d$ at time $t$ as $d_t$, and we define several features based on each paper's topic distribution, including popularity, novelty, diversity, and authority. We provide details on these features next.

First, we consider that as popular topics tend to attract more attention and resources than relatively unpopular ones, it is relatively easy for papers related to such topics to accrue citations. To capture this effect, we quantify the popularity of each topic $z$ across the overall corpus by $popularity(z) = \sum_{d \in C} p(z|d) \times c_d$, where $p(z|d)$ is the probability that paper $d$ distributes on topic $z$ and $c_d$ is the number of citations $d$ collects until the timestamp $t$. The popularity of a target paper $d_t$ (paper $d$ at time $t$) is then defined as:

$$C\text{-}popularity(d_t) = \sum_{z \in Z} popularity(z) \times p(z|d_t). \tag{6.1}$$

Second, a paper's novelty is an essential factor when assessing its contribution to the scientific community. Previous work assumes that the novelty of an article can be determined by measuring the difference between its content and that of its references [234]. We utilize the Kullback-Leibler divergence [113] to capture the sum of the difference between $d_t$'s topic distribution and the topic distribution of each of

its references. Specifically, we define the novelty of paper $d_t$ as

$$\textit{C-novelty}(d_t) = \frac{\sum_{d_r \in R} KL(p(Z|d_t), p(Z|d_r))}{|R|}, \tag{6.2}$$

where $KL(p(Z|d_t), p(Z|d_r)) = \sum_{z \in Z} \log \frac{p(z|d_t)}{p(z|d_r)} p(z|d_t)$ and $R$ is the set of $d_t$'s references.

Third, the topic diversity of a paper, defined as the breadth of its topic distribution, is able to distinguish between different types of papers, such as surveys and technical work. We follow the definition of diversity in [234] as calculated by Shannon entropy:

$$\textit{C-diversity}(d_t) = \sum_{z \in Z} -p(z|d_t) \log p(z|d_t). \tag{6.3}$$

Fourth, Kleinberg has pointed out that in a hyperlinked web environment, a "good" authority represents a page that is linked to by many hubs [106]. Similarly, academic authority can be designated by being highly cited by many other researchers in a specific domain of expertise. To measure the authority of researcher $a$ on topic $z$, we propose the following definition: $\textit{authority}(a, z) = \sum_{d \in C_a} p(z|d) \times c_d$, where $C_a$ is the researcher $a$'s previous publications. Therefore, given the target paper $d_t$, the author's authority is distributed over the topic distribution of $d_t$. Formally,

$$\textit{C-authority}(d_t, a) = \sum_{z \in Z} p(z|d_t) \times \textit{authority}(a, z). \tag{6.4}$$

This definition of authority follows from the intuition that a correspondence between a paper's topic distribution and its authors' expertise can help ensure its quality.

**Venue Factors.** Top venues attract high-quality submissions, and high-quality

submissions elevate the reputation of their respective venues. Google Scholar metrics show that different venues have large differences in their $h5$-indices (the $h$-index computed only from articles published within the last 5 complete years), a measure of venue impact. For example, in the field of data mining and analysis, the top three venues are ACM SIGKDD, IEEE TKDE, and ACM WSDM, with $h5$-indices of 69, 57, and 54, respectively. By contrast, most other venues in this field typically have $h5$-indices between 10 and 20. In light of these differences, we engage in the investigation of how different venues influence the probability that a paper contributes to its author's $h$-index. Two heuristic metrics are examined, namely (1) the average number of citations each paper in the venue collects and (2) the ratio between the number of papers in the venue with at least max-$h$-index citations to the venue's total number of papers. Every researcher aims to publish scientific results in well-respected journals and conferences, so our intuition is that top venues help researchers spread their scientific impact and, more specifically, to increase the citation counts of their papers, which further offers a potential to increase their $h$-indices.

**Social Factors.** Previous studies have shown that researchers display a tendency to cite their co-authors' work [16]. As shown in Figure 6.5(c), our investigations reveal that a researcher's $h$-index is also positively correlated with his or her total number of collaborators. To explore this trend, we extract a weighted collaboration network from the dataset, where each author is denoted as a node and each link between two nodes is connected if the researchers have collaborated with each other. The weight of each link is defined as the frequency of collaboration. We then extract four features for each node (author) from the collaboration network, including the number of co-authors (degree), the PageRank score, the average $h$-index of co-authors, and the weighted average $h$-index of co-authors. For a given paper, the highest values among its authors for these four metrics are used as social factors.

**Reference Factors.** The scientific impact of a scholarly work is often quantified

TABLE 6.4

EXISTING FACTOR DEFINITIONS AND CORRELATIONS

| Factor | Description | $P_{new}^{max}$ | | $P_{new}^{first}$ | |
|---|---|---|---|---|---|
| | | $cc_{2002}$ | $cc_{2007}$ | $cc_{2002}$ | $cc_{2007}$ |
| *E-numc* | #citations | 0.1656 | 0.2352 | 0.1509 | 0.2029 |
| *E-numc-ave* | #ave-c per year | 0.1913 | 0.3203 | 0.1579 | 0.2600 |
| *E-num-years* | #publication-years | 0.0140 | 0.0856 | 0.0103 | 0.0415 |

by its respective citation count. The more times a publication is cited by others, the greater its assumed impact. Conversely, as most scientific research is undertaken by "standing on the shoulder of giants," we ask whether highly cited papers actually tend to acknowledge the previous "giants" upon whom they stand. Two intuitive factors are used to evaluate this question, namely (1) the ratio of a paper's references that have at least max-$h$-index citations to the paper's total number of references and (2) the average number of citations accumulated by the paper's references.

**Temporal Factors.** Just as fast-rising phenomena typically attract the attention of crowds more easily, a "rising star" in academia can attract wide publicity. Previous work has found that temporal information can be a powerful factor in modeling scientific impact [16, 234], so it is straightforward to assume that the speed at which an author's $h$-index grows should affect the rate at which the author's papers contribute to his or her $h$-index. To capture this effect, we examine the increase of authors' $h$-indices within the past three years. Specifically, we consider four temporal factors, including the $h$-index changes of the first author, the max-$h$-index author, and the average change and maximum change among all authors. The specific definitions are shown in Table 6.2.

### 6.5.3 Existing Factors for Previous Papers

Besides the above-examined factors, which generally drive papers to increase authors' $h$-indices, we discuss several factors that are extracted from the existing citation information for papers published before time $t$. For each paper, we consider three intuitive factors: (1) the total number of citations the paper has accrued until $t$; (2) the average number of citations the paper has accrued per year until $t$; and (3) the length of time between the paper's publication date and $t$.

The correlation of each factor with the target variable is provided in Table 6.4. We observe that, from among these factors, the average number of citations per year that each paper has accrued before $t$ is most highly correlated with the probability that the paper will increase its primary author's future $h$-index at time $t + \Delta t$.

### 6.5.4 Summary

Drawn from the correlation analysis above, we provide the following intuitions relating to academia:

First, a research scholar's future $h$-index is highly correlated with his or her current impact—namely, the researcher's $h$-index—rather than the number of citations each of his or her publications collect or the length of his or her academic career.

Second, a scientific researcher's authority on a topic is the most decisive factor in facilitating an increase in his or her $h$-index. This coincides with the fact that the society fellows or lifetime honors are typically conferred for contributions to a particular topic or domain. However, the topic diversity of a publication is a relatively non-effective factor in growing its scientific impact, as measured by its probability to increase the $h$-indices of its authors.

Third, the reputation of the venue in which a given paper is published is another crucial factor in determining the probability that it will contribute to its authors' $h$-indices. Top venues distinguish one's work as outstanding and expand one's scientific

impact; gradually, one's impact can further help to increase the venue's prestige.

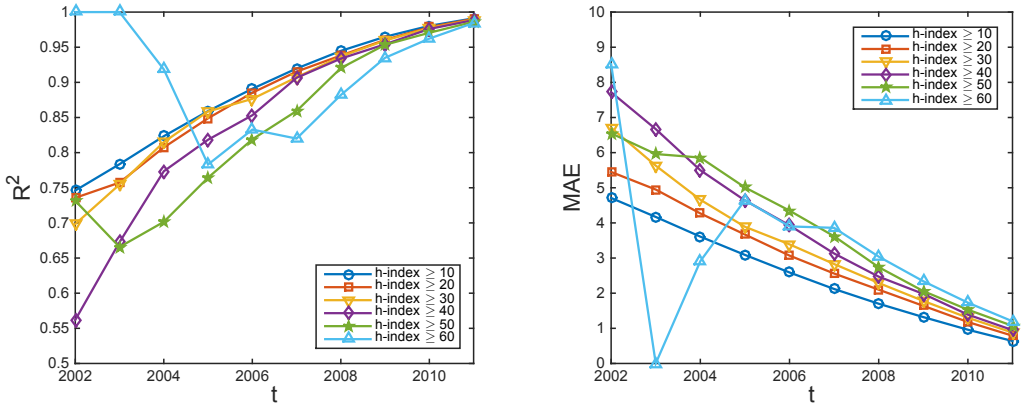
Finally, while people in social society often follow vogue trends, publishing on an academically "hot" but unfamiliar topic is unlikely to further one's scientific impact, at least as measured by one's $h$-index.

## 6.6 Scientific Impact Prediction

In this section, we demonstrate the predictability of scientific impact in two parts. First, we predict the future $h$-indices of scientific scholars. Second, given the estimated future $h$-indices, we determine whether a previously ($P_{old}$) or newly ($P_{new}$) published paper will contribute to its primary author's $h$-index within a given timeframe.

### 6.6.1 Experimental Setup

Our primary task is to predict whether a paper published by (at or before) timestamp $t$ will contribute to the future $h$-index of its primary author within a given time period $\Delta t$. To accomplish this, we need to first estimate the author's $h$-index at $t + \Delta t$ based on data observed at time $t$. For example, by setting $t = 2007$, $\Delta t = 5$ years, and the minimum $h$-index of the primary author to 10, we collect one set of papers ($P_{new}$) published in 2007 and another set of papers published before 2007 ($P_{old}$). We then extract the features from the corpus observed at 2007 and observe whether the number of citations for each paper in these two sets is larger than or equal to the future $h$-index of its primary author in 2012 (the last year represented in our dataset).




(a) Predictive performance ($R^2$)

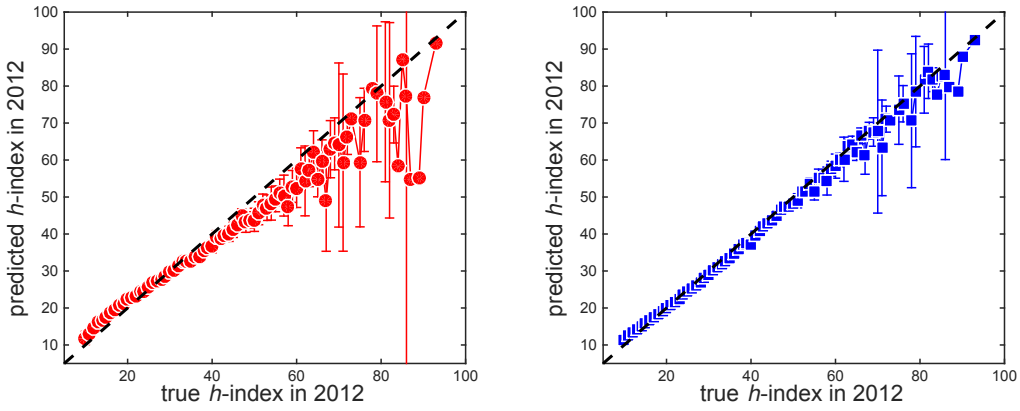
(b) Predictive performance (MAE)

Figure 6.7. Performance for predicting future $h$-indices.

### 6.6.2 Predicting Future $h$-indices

**Methods.** Similar to the previous work of [2], wherein Acuna et al. propose a method to infer the future $h$-indices of neuroscientists, our $h$-index prediction problem is formulated as a regression task. For this task, we use linear regression, primarily due to its effectiveness, simplicity, and interpretability. The features used here contain the factors detailed in Table 6.1. To quantitatively evaluate the model predictions, we report the performance in terms of the coefficient of determination ($R^2$) [133] and the mean absolute error (MAE).

**Prediction Results.** We present the extent to which research scholars' future $h$-indices can be inferred from their previous publication records. Figure 6.7 reports the predictive performance in terms of $R^2$ and MAE. On the one hand, the rising lines in Figure 6.7(a) and the descending lines in Figure 6.7(b) as $t$ increases both imply that our prediction task is easier when given a shorter timeframe. That is, future $h$-indices are more predictable when the future is close to $t$. Our observations agree with the intuition that the variability in the development of researchers' $h$-indices increases with a large prediction timeframe. On the other hand, the figure generally




(a) Prediction from 2002 to 2012 (b) Prediction from 2007 to 2012

Figure 6.8. $h$-indices in data vs. predicted $h$-indices.

suggests that our prediction task is more difficult for authors with high $h$-indices. Intuitively, as an author's $h$-index increases, the variability in the development of his or her scientific impact also increases, which results in an increasingly challenging prediction task.

Figure 6.8 illustrates the concordance between the future $h$-indices predicted by our model and the actual $h$-indices according to the provided data. As the prediction timeframe can be varied, Figure 6.8(a) reports results over a ten-year timeframe, while Figure 6.8(b) reports results over a five-year timeframe. For both plots, optimal performance is denoted by the dashed $y = x$ line, which represents perfect agreement between the predictions and data. From the plots we observe that higher $h$-indices correspond to higher variability (error bars) and increasing deviation from optimal performance, suggesting that higher future $h$-indices are more difficult to predict. However, Figure 6.8(a) also demonstrates higher levels of deviation and variability than Figure 6.8(b), indicating that accurately predicting future $h$-indices is more difficult over longer timeframes.

### 6.6.3 Predicting Whether Papers Increase $h$-indices

**Methods.** Our problem of predicting whether a paper increases its primary author's future $h$-index is formulated as a classification task. For this task, we employ a series of standard classification models, including logistic regression (LRC), support vector machine (SVM), naïve Bayes (NB), radial basis function network (RBF), bagged decision trees (BAG), and random forest (RF). Generally, we report the prediction results of each method to demonstrate the predictability of scientific impact, though we only use logistic regression to analyze factor contributions and parameter settings.

Recall that for this task, we have defined two sets of papers, $P_{new}$ and $P_{old}$, and we generate predictions for both. When defining the primary author as either the max-$h$-index author or the first author, we further extract two sets of papers from both $P_{new}$ and $P_{old}$, respectively, and have $P_{new}^{max}$, $P_{new}^{first}$, $P_{old}^{max}$, and $P_{old}^{first}$. For each set of papers, we use half of the instances (papers) in the set for model training and the remaining half for model validation. When predicting for $P_{new}$, we use the six groups of 24 total factors described in Table 6.2. When predicting for $P_{old}$, these 24 factors are used along with the three additional factors described in Table 6.4. To quantitatively evaluate the predictability of the problem, we repeat the prediction experiments ten times and report the average performance in terms of precision, recall, $F_1$ score, area under the receiver operating characteristic (AUC), and accuracy. Furthermore, as our problem can be viewed as a ranking task (i.e., rank all of a scholar's papers in the reverse order of probability that they will increase her/his $h$-index), the precision at the top 3 results (Pre@3) and mean average precision (MAP) are also used to evaluate performance.

TABLE 6.5

PREDICTIVE PERFORMANCE FOR $P_{new}^{max}$

| | Method | Precision | Recall | $F_1$ | AUC | Accuracy | Pre@3 | MAP |
|---|---|---|---|---|---|---|---|---|
| | Random | 0.2107 | 0.5000 | 0.2965 | 0.5000 | 0.5000 | 0.5899 | 0.4132 |
| | LRC | 0.8233 (0.0049) | 0.5929 (0.0062) | 0.6894 (0.0038) | 0.9299 (0.0017) | 0.8873 (0.0010) | 0.8928 | 0.9440 |
| | SVM | 0.8377 (0.0050) | 0.5806 (0.0044) | 0.6858 (0.0034) | 0.7753 (0.0021) | 0.8879 (0.0011) | 0.8033 | 0.8655 |
| $P_{new}^{max}$ | NB | 0.6483 (0.0113) | 0.5371 (0.0151) | 0.5873 (0.0072) | 0.8497 (0.0043) | 0.8409 (0.0024) | 0.8201 | 0.8759 |
| | RBF | 0.6679 (0.0109) | 0.5573 (0.0124) | 0.6075 (0.0081) | 0.8403 (0.0078) | 0.8482 (0.0029) | 0.7897 | 0.8694 |
| | BAG | 0.7992 (0.0045) | 0.7455 (0.0111) | 0.7713 (0.0043) | 0.9548 (0.0008) | 0.9068 (0.0009) | 0.8919 | 0.9509 |
| | RF | 0.7647 (0.0058) | 0.7630 (0.0090) | 0.7638 (0.0043) | 0.9373 (0.0015) | 0.9005 (0.0016) | 0.8734 | 0.9376 |

TABLE 6.6

PREDICTIVE PERFORMANCE FOR $P_{new}^{first}$

| | Method | Precision | Recall | $F_1$ | AUC | Accuracy | Pre@3 | MAP |
|---|---|---|---|---|---|---|---|---|
| | Random | 0.2660 | 0.5000 | 0.3472 | 0.5000 | 0.5000 | 0.8068 | 0.6728 |
| | LRC | 0.8202 (0.0106) | 0.6129 (0.0131) | 0.7014 (0.0077) | 0.9112 (0.0028) | 0.8611 (0.0027) | 0.9200 | 0.9647 |
| | SVM | 0.7866 (0.0207) | 0.4893 (0.0134) | 0.6031 (0.0114) | 0.7205 (0.0065) | 0.8059 (0.0048) | 0.8666 | 0.9094 |
| $P_{new}^{first}$ | NB | 0.6776 (0.0149) | 0.5176 (0.0234) | 0.5865 (0.0143) | 0.8316 (0.0064) | 0.8130 (0.0046) | 0.8733 | 0.9250 |
| | RBF | 0.6895 (0.0167) | 0.5418 (0.0252) | 0.6064 (0.0163) | 0.8200 (0.0059) | 0.8661 (0.0057) | 0.8866 | 0.9277 |
| | BAG | 0.7815 (0.0103) | 0.6901 (0.0092) | 0.7329 (0.0068) | 0.9216 (0.0023) | 0.8661 (0.0035) | 0.9000 | 0.9609 |
| | RF | 0.7322 (0.0139) | 0.7136 (0.0131) | 0.7227 (0.0111) | 0.9033 (0.0034) | 0.8542 (0.0060) | 0.8800 | 0.9518 |

TABLE 6.7

PREDICTIVE PERFORMANCE FOR $P_{old}^{max}$

| | Method | Precision | Recall | $F_1$ | AUC | Accuracy | Pre@3 | MAP |
|---|---|---|---|---|---|---|---|---|
| | Random | 0.3776 | 0.5000 | 0.4303 | 0.5000 | 0.5000 | 0.5070 | 0.3186 |
| | LRC | 0.9840 (0.0006) | 0.9829 (0.0008) | 0.9834 (0.0004) | 0.9995 (0.0000) | 0.9874 (0.0003) | 0.9992 | 0.9992 |
| | SVM | 0.9835 (0.0009) | 0.9806 (0.0014) | 0.9820 (0.0008) | 0.9853 (0.0007) | 0.9864 (0.0005) | 0.9825 | 0.9844 |
| $P_{old}^{max}$ | NB | 0.9316 (0.0024) | 0.8290 (0.0040) | 0.8773 (0.0022) | 0.9763 (0.0008) | 0.9124 (0.0014) | 0.9620 | 0.9601 |
| | RBF | 0.7860 (0.1066) | 0.6965 (0.1440) | 0.7211 (0.0533) | 0.8768 (0.0060) | 0.8019 (0.0181) | 0.8933 | 0.8902 |
| | BAG | 0.9939 (0.0005) | 0.9898 (0.0003) | 0.9918 (0.0003) | 0.9998 (0.0000) | 0.9938 (0.0002) | 0.9998 | 0.9997 |
| | RF | 0.9816 (0.0020) | 0.9880 (0.0003) | 0.9848 (0.0011) | 0.9992 (0.0001) | 0.9884 (0.0008) | 0.9984 | 0.9984 |

TABLE 6.8

PREDICTIVE PERFORMANCE FOR $P_{old}^{first}$

| | Method | Precision | Recall | $F_1$ | AUC | Accuracy | Pre@3 | MAP |
|---|---|---|---|---|---|---|---|---|
| $P_{old}^{first}$ | Random | 0.4753 | 0.5000 | 0.4873 | 0.5000 | 0.5000 | 0.6424 | 0.4524 |
| | LRC | 0.9818 (0.0011) | 0.9803 (0.0007) | 0.9810 (0.0004) | 0.9988 (0.0000) | 0.9819 (0.0003) | 0.9990 | 0.9994 |
| | SVM | 0.9838 (0.0056) | 0.9725 (0.0085) | 0.9781 (0.0024) | 0.9790 (0.0024) | 0.9792 (0.0021) | 0.9827 | 0.9865 |
| | NB | 0.9588 (0.0030) | 0.7963 (0.0051) | 0.8700 (0.0024) | 0.9713 (0.0009) | 0.8868 (0.0017) | 0.9740 | 0.9814 |
| | RBF | 0.8956 (0.0244) | 0.4829 (0.0505) | 0.6259 (0.0428) | 0.8288 (0.0226) | 0.7271 (0.0218) | 0.8810 | 0.8932 |
| | BAG | 0.9873 (0.0010) | 0.9842 (0.0009) | 0.9858 (0.0004) | 0.9993 (0.0001) | 0.9865 (0.0003) | 0.9990 | 0.9993 |
| | RF | 0.9762 (0.0024) | 0.9828 (0.0009) | 0.9795 (0.0013) | 0.9982 (0.0002) | 0.9804 (0.0012) | 0.9975 | 0.9985 |

**Prediction Results for $P_{new}$.** The predictability of whether a paper published at $t = 2007$ will contribute to its primary author's future $h$-index within $\Delta t = 5$ years is presented in Tables 6.5 and 6.6. The prediction is applied to the papers whose primary author had an $h$-index in 2007 of at least 10. The resulting set when considering the max-$h$-index author as the primary author, $P_{new}^{max}$, contains 29,254 papers, of which 21.07% successfully contributed to their primary author's future $h$-index by 2012. When the first author serves the primary author, the resulting set $P_{new}^{first}$ covers 9,231 papers, of which 26.60% increased the first author's future $h$-index by 2012.

Overall, when predicting $P_{new}^{max}$, random guessing achieves an $F_1$ score of 0.2965 and an accuracy of 0.5000. However, our methodology achieves a predictive power that significantly outperforms random guessing, demonstrating an $F_1$ score ranging from 0.5873 to 0.7713 (+98% to +160% increase) and an accuracy ranging from 0.7753 to 0.9548 (+66% to +91% increase). The performance is similarly promising when measured by precision, recall, and AUC. Furthermore, by ranking all of a scholar's publications in the reverse order of probability that they increase his or her $h$-index, logistic regression can achieve a Pre@3 of 0.8928 and a MAP of 0.9440. Similarly, the experimental performance when predicting for $P_{new}^{first}$, where the first author is considered the primary author, significantly outperforms random guessing and demonstrates a comparable predictability with the results for $P_{new}^{max}$.

**Prediction Results for $P_{old}$.** The predictability of whether a paper published before $t = 2007$ will contribute to its primary author's future $h$-index ($\geq 10$) within $\Delta t = 5$ years is presented in Tables 6.7 and 6.8. The resulting set when considering the max-$h$-index author (the first author) as the primary author, $P_{old}^{max}$ ($P_{old}^{first}$), contains 161,348 (85,704) papers, of which 37.76% (47.53%) successfully contributed to their primary authors' future $h$-indices by 2012. Random guessing achieves an $F_1$ score of 0.4303 (0.4873), an AUC of 0.5000 (0.5000), and a Pre@3 of 0.5070

(0.6424). Generally, the algorithms can achieve at least twice the performance of random guessing, as measured by all of the evaluation metrics employed. The results demonstrate strong predictability for this scientific impact prediction task, with performance scores ranging from 0.98–0.99 as measured by precision, recall, $F_1$ score, AUC, accuracy, Pre@3, and MAP.

As the selected algorithms achieve similarly effective results, we use logistic regression to examine the remaining experiments—primarily owing to its interpretability.

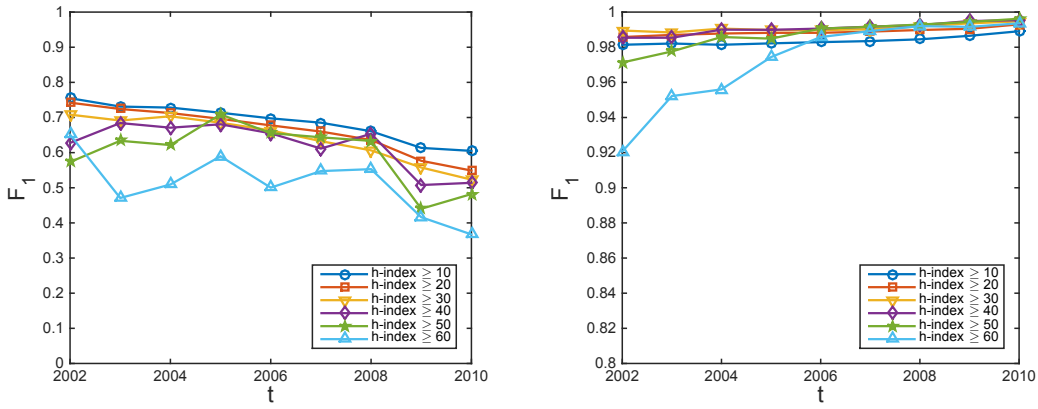
### 6.6.4 Predictability of Different Papers

Our experimental results provide evidence for the predictability of whether a newly or previously published paper will contribute to the $h$-index of its primary author within five years. Yet, two intuitive questions naturally arise concerning this predictability: First, is a primary author with a high or low $h$-index more predictable? Second, is a paper more predictable given a long or short prediction timeframe?

To answer these questions, we investigate the predictability of papers conditioned on the primary author's $h$-index and the length of the given prediction timeframe ($\Delta t$). Figure 6.9 shows the predictive performance given different constraints for four sets of papers, conditioned on the publication date and primary author definition—$P_{new}^{max}$, $P_{old}^{max}$, $P_{new}^{first}$, and $P_{old}^{first}$.

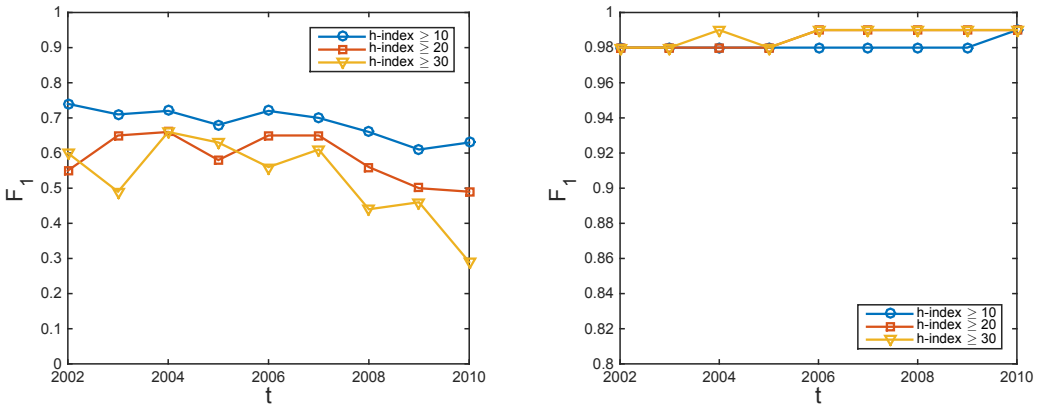
First, from Figures 6.9(a) and 6.9(c), we find that predicting for papers with low-$h$-index primary authors is a relatively easy task as measured by $F_1$ vis-à-vis predicting for those with high $h$-indices.

Intuitively, the prediction task becomes increasingly non-trivial because of the increasing difficulty for any particular paper to reach the defined local threshold (i.e., the primary author's $h$-index). Additionally, we observe that performance generally decreases as $t$ increases, implying that our prediction task is easier when given a longer timeframe $\Delta t = 2012 - t$. Intuitively, papers can accrue more citations as time




(a) Predicting for $P_{new}^{max}$

(b) Predicting for $P_{old}^{max}$



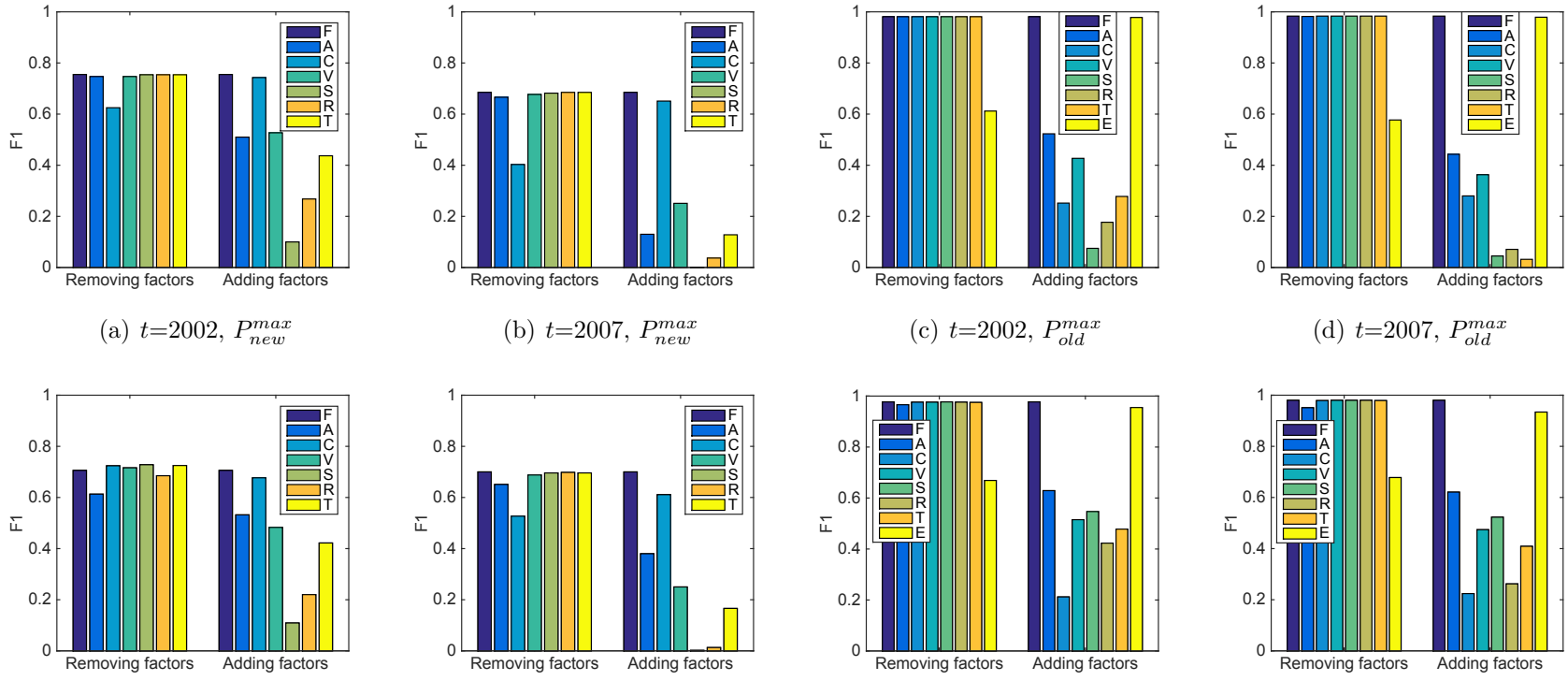

(c) Predicting for $P_{new}^{first}$

(d) Predicting for $P_{old}^{first}$

Figure 6.9. Predictive performance for different papers.

goes on, during which time the authors' influence may increase, which may further compound the rate at which citations accrue. In summary, determining which newly published papers will increase one's $h$-index is more predictable when conducted over a relatively long timeframe for those who have relatively low $h$-indices.

Note that from Figures 6.9(b) and 6.9(d), we can see that when predicting for previously published papers, both observations above are not significant. This is due to the relatively strong predictability of those papers.




(a) $t$=2002, $P_{new}^{max}$ (b) $t$=2007, $P_{new}^{max}$ (c) $t$=2002, $P_{old}^{max}$ (d) $t$=2007, $P_{old}^{max}$

(e) $t$=2002, $P_{new}^{first}$ (f) $t$=2007, $P_{new}^{first}$ (g) $t$=2002, $P_{old}^{first}$ (h) $t$=2007, $P_{old}^{first}$

Figure 6.10. Factor contribution analysis. Logistic regression model trained with only or without the denoted factors. F: full feature set; A: Author factors; C: Content factors; V: Venue factors; S: Social factors; R: Reference factors; T: Temporal factors; E: Existing factors for previously published papers. The left and right sides of the figure illustrate the effects of omitting (the "without" case) and only including (the "with only" case) the indicated group of factors for model training, respectively.

TABLE 6.9

INFORMATION GAIN RATIO (IGR) OF EACH FACTOR

| Factor | $P_{new}^{max}$ | | $P_{old}^{max}$ | |
|---|---|---|---|---|
| | $\text{IGR}_{2002}$ (R) | $\text{IGR}_{2007}$ (R) | $\text{IGR}_{2002}$ (R) | $\text{IGR}_{2007}$ (R) |
| *A-first-max* | 0.0193 (15) | 0.0255 (10) | 0.0168 (10) | 0.0206 (10) |
| *A-ave-max* | 0.0126 (19) | 0.0200 (11) | 0.0153 (11) | 0.0207 (9) |
| *A-sum-max* | 0.0229 (13) | 0.0193 (12) | 0.0170 (9) | 0.0134 (11) |
| *A-first-ratio* | 0.0133 (17) | 0.0111 (15) | 0.0138 (12) | 0.0114 (12) |
| *A-max-ratio* | 0.0631 (5) | 0.0409 (7) | 0.0665 (7) | 0.0656 (7) |
| *A-num-authors* | 0.0079 (20) | 0.0044 (23) | 0.0025 (21) | 0.0007 (26) |
| *C-popularity* | 0.0315 (11) | 0.0053 (20) | 0.0024 (23) | 0.0035 (23) |
| *C-diversity* | 0.0258 (12) | 0.0047 (22) | 0.0018 (26) | 0.0031 (25) |
| *C-novelty* | 0.0127 (18) | 0.0062 (19) | 0.0018 (25) | 0.0000 (27) |
| *C-auth.-first* | 0.3988 (1) | 0.3407 (2) | 0.0858 (3) | 0.1269 (4) |
| *C-auth.-max* | 0.3006 (3) | 0.2651 (3) | 0.0678 (6) | 0.1081 (5) |
| *C-auth.-ave* | 0.3781 (2) | 0.3462 (1) | 0.0854 (4) | 0.1327 (3) |
| *V-h-index* | 0.0619 (6) | 0.0714 (5) | 0.0494 (8) | 0.0586 (8) |
| *V-citation* | 0.1233 (4) | 0.1090 (4) | 0.0845 (5) | 0.1009 (6) |
| *S-degree* | 0.0000 (24) | 0.0029 (24) | 0.0018 (24) | 0.0071 (19) |
| *S-pagerank* | 0.0000 (23) | 0.0052 (21) | 0.0025 (22) | 0.0089 (16) |
| *S-h-coauthor* | 0.0065 (21) | 0.0091 (17) | 0.0077 (21) | 0.0076 (17) |
| *S-h-weight* | 0.0045 (22) | 0.0078 (18) | 0.0051 (20) | 0.0056 (21) |

TABLE 6.9

*Continued*

| Factor | $P_{new}^{max}$ | | $P_{old}^{max}$ | |
|---|---|---|---|---|
| | $\text{IGR}_{2002}$ (R) | $\text{IGR}_{2007}$ (R) | $\text{IGR}_{2002}$ (R) | $\text{IGR}_{2007}$ (R) |
| *R-h-index* | 0.0180 (16) | 0.0167 (14) | 0.0104 (16) | 0.0111 (14) |
| *R-citation* | 0.0196 (14) | 0.0096 (16) | 0.0110 (14) | 0.0113 (13) |
| *T-ave-h* | 0.0551 (7) | 0.0506 (6) | 0.0104 (17) | 0.0058 (20) |
| *T-max-h* | 0.0476 (8) | 0.0291 (9) | 0.0113 (13) | 0.0041 (22) |
| *T-h-first* | 0.0370 (9) | 0.0386 (8) | 0.0108 (15) | 0.0072 (18) |
| *T-h-max* | 0.0341 (10) | 0.0168 (13) | 0.0093 (18) | 0.0034 (24) |
| *E-numc* | \ | \ | 0.7324 (2) | 0.7598 (1) |
| *E-numc-ave* | \ | \ | 0.7336 (1) | 0.6477 (2) |
| *E-num-years* | \ | \ | 0.0002 (27) | 0.0105 (15) |

### 6.6.5 Factor Contribution Analysis

To predict whether a paper will increase its primary author's $h$-index, we devise six diverse groups of factors (see §6.5) that may drive the growth of scientific impact.

To explore the contributions and importance of each factor group to the prediction task, we employ a "jackknife" approach with two cases: (1) one at a time, we remove a group of factors and evaluate the predictive performance of our model trained only on the remaining five groups (the "without" case); and (2) one at a time, we use only a single group of factors and evaluate the predictive performance of our model trained only on this group (the "with only" case). This approach provides information on the individual contribution and unique information that each group of factors supplies

to the overall prediction task. Figure 6.10 provides the $F_1$ scores for the two cases with different $t$ (2002 and 2007), primary authors (max-$h$-index and first authors), and publication dates (new and old). We can see that the contributions of different groups of factors demonstrate a high degree of variability.

In Figures 6.10(a) and 6.10(b), the ∼20% drop in $F_1$ score demonstrated by the removal of content factors indicates that they are critically important to predicting for $P_{new}^{max}$. By contrast, the marginal decreases in performance demonstrated by the removal of other types of factors imply that the remaining factors provide a limited amount of unique information. When used only by themselves, the content factors still play the most important role in predicting the growth of scientific impact, though venue factors also achieve a marked effect on performance. Furthermore, with the exclusion of content factors, all groups of factors demonstrate greater importance when employed over a longer timeframe $\Delta t$.

From Figures 6.10(c) and 6.10(d), we can see that the existing factors are crucially important to predicting for $P_{old}^{max}$, both by themselves (the "with only" contributions) and when used with other factors (the "without" contributions). Different from predicting for $P_{new}^{max}$ in Figures 6.10(a) and 6.10(b), author factors play a more important role than both content and venue factors, observed from the "with only" factor contributions. Overall, we find that this contribution analysis is consistent with the factor correlation results elaborated upon in the previous section.

Figures 6.10(e) and 6.10(f) show that when predicting for newly published papers, the content, author, and venue factors contribute the most to the increase of the first authors' future $h$-indices. Similarly, from Figures 6.10(g) and 6.10(h), we can see that the existing information before $t$ is the most decisive factor group for predicting whether the previously published papers can contribute to the first authors' future $h$-indices. Surprisingly, we also find that different from the prediction cases in $P_{new}^{max}$, $P_{old}^{max}$, and $P_{new}^{first}$, the role of social factors is comparable with author and venue

factors when predicting for $P_{old}^{first}$.

In summary, when predicting for the newly published papers in Figures 6.10(a), 6.10(b), 6.10(e) and 6.10(f), the content factor group is most crucial to generating effective predictions, followed by venue, author, and temporal factors. However, observed from Figures 6.10(c), 6.10(d), 6.10(g) and 6.10(h), the existing factor group is the most telling followed by author and venue factor groups when predicting for previously published papers. The group of content factors is important when predicting for the increase of the max-$h$-index authors, while its effect is not significant compared to other factors when predicting the contribution to the first authors' $h$-indices.

We further examine the contributions of each individual factor to the prediction tasks. To assess each factor's importance, we employ the measure of information gain ratio (IGR) [113], which is based on the expected reduction in entropy—that is, uncertainty—achieved by learning the state of a given factor. The higher the IGR for a given factor, the greater its measured importance.

Table 6.9 lists the IGR and corresponding ranking for each individual factor. When considering the IGR for $P_{new}$, the factors that are indicative of an author's topical authority are the most important, including *C-authority-max*, *C-authority-ave* and *C-authority-first*. Following in importance are the two venue factors. When considering the IGR for $P_{old}$, the factors that are indicative of the number of existing citations (*E-numc* and *E-numc-ave*) achieve the top two positions, followed by author authorities and venue factors. The IGR calculated for the remaining factors decreases to the next lowest order of magnitude, indicating that they provide relatively limited contributions to our prediction tasks.

### 6.6.6 Prototype $h$-index Prediction Tool

In light of our investigations into the factors that influence authors' $h$-indices, we have developed an online tool that allows users to generate $h$-index predictions based

on our findings. An image of the working prototype is provided as Figure 6.11.

The tool provides separate functionality for predicting the development of authors' $h$-indices (left) and predicting whether a paper will contribute to its authors' $h$-indices (right). To predict the development of authors' $h$-indices, users may enter basic author details, such as an author's current $h$-index, number of publications, and initial year of publication. To predict the probability that a paper will contribute to its authors' $h$-indices, users may enter basic paper details, such as the title, author list, year, venue, and abstract text. These details are then used to generate the factors described in this work, which serve as input to the $h$-index growth or paper contribution model developed through our investigations.

We hope that the tool may be used by scholars to more effectively disseminate their work and to better gauge their future scientific impact.

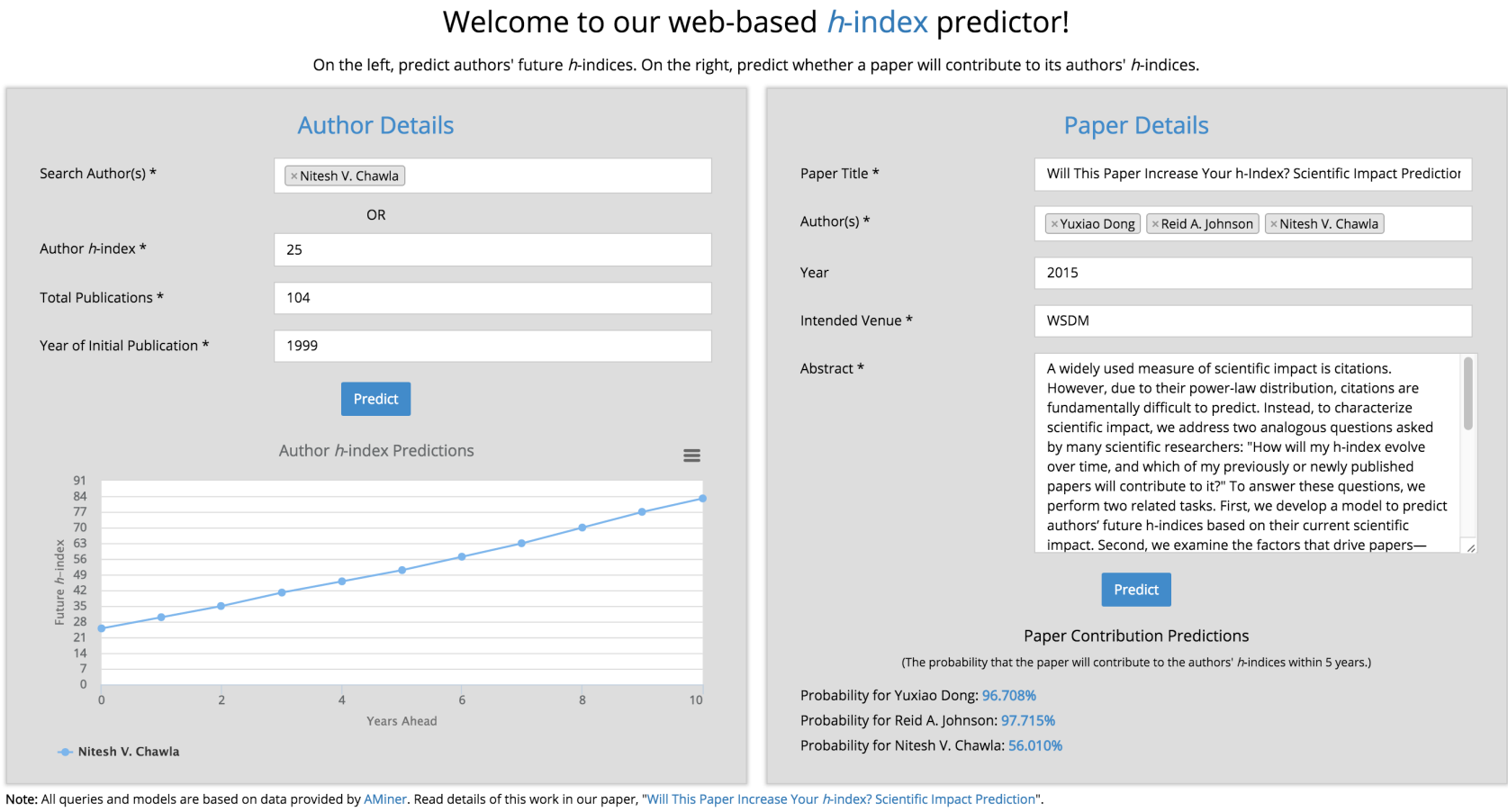



Figure 6.11. Prototype $h$-index prediction tool (see `http://www.icensa.com/hindex`). The prototype provides two distinct functionalities. On the left, the tool can be used to provide predictions of the development of authors' $h$-indices. On the right, the tool can be used to predict whether a paper will contribute to its authors' $h$-indices.

## 6.7 Related Work

Scientific impact modeling is being extensively explored and has become an important and popular research topic [2, 39, 46, 220, 224]. Its study offers the potential to help scholars more effectively disseminate their work and expand their scientific influence.

Traditionally, the number of citations has been widely used as a measurement of scientific impact for both individual papers and solitary scientific researchers. Several practical metrics have been designed to reflect scientific impact based on citations. Garfield proposed the impact factor for indexing and evaluating the quality of journals [69]. More recently, Hirsch proposed the $h$-index, which attempts to measure both a researcher's productivity and the popularity of his or her published work [90]. Both impact factor and $h$-index successfully characterize the motivations and behavior of the scientific community, where scholars aspire to publish results in high-impact venues to increase their influence and $h$-indices and venues aim to publish cogent, influential work to improve their reputations and impact factors.

Besides its measurement, a large body of work has been focused on the prediction of scientific impact. The 2003 ACM SIGKDD Cup introduced a competition focused around citation count prediction [70], with the task of estimating the number of times a paper has been cited given its previous number of citations. Following this, many efforts have been made to predict the number of future citations for scholarly work. Castillo et al. studied the correlation between author reputation and citations [28]. Yan et al. examined a series of features important to future citations [233, 234]. Wang et al. uncovered basic mechanisms that govern scientific impact, which has the power to quantify and predict citation counts [181, 224]. However, the effectiveness of such predictions is fundamentally limited by the heavy-tailed distribution of citations.

Herein we (re)define the impact prediction problem by addressing a related question, namely: "which of my papers will increase my (future) $h$-index?" The crucial

difference between ours and previous work is that rather than trying to solve a regression task in a highly skewed environment, we instead tackle the problem by generating a local threshold (the author's $h$-index) for each paper's future citation count.

Our work is also related to other mining tasks in academic data such as citation pattern and recommendation [40, 174, 186, 202, 237], topic influence [126, 204], information flow [184, 185], collaboration prediction [199, 222], and analysis of citation networks [221] and academic social networks [203]. Further, as the formalization of our predictive task is partly inspired by the cascade growth prediction problem [32], the prediction of scientific impact is related to predicting the popularity [5, 93, 170] of online "paper" (e.g., tweet, video, photo) in social media.

## 6.8 Conclusion

In this chapter, we study the predictability of scientific impact by formalizing two problems that can be reduced to the following questions: How will my $h$-index evolve over time, and which of my papers will contribute to it? Our primary task is to determine whether a given paper, either previously or newly published, will increase the *future* $h$-index of its primary author within a predefined timeframe. To address this task, we first formalize an $h$-index prediction problem to estimate researchers' future $h$-indices. We then use these estimates as the target for prediction in our primary task, which offers a powerful way of quantifying the interplay between researchers and publications and their effects on scientific impact.

Surprisingly, we find that topic diversity and popularity have no statistical correlation with whether a paper will contribute to its primary author's future $h$-index. We also find that two factors—topical authority and publication venue—are critical in determining whether a newly published paper will contribute to its primary author's future $h$-index, while the existing citation count is the most decisive factor for a previously published paper. We demonstrate that the contribution of a paper to

the impact of a researcher with a higher $h$-index is generally more difficult to predict than for a researcher with a lower $h$-index. Finally, we develop an $h$-index prediction tool informed by our findings. Overall, our work demonstrates a greater than 90% potential predictability, as measured by accuracy, for whether a paper will contribute to its primary author's $h$-index within five years.

Future work could study the interplay between a researcher's estimated future $h$-index and the set of papers that we predict will contribute to his or her $h$-index. Furthermore, as this work is conducted only on literature from computer science, examining other scientific disciplines for the same observed patterns could widen the scope and significance of our findings.

# CHAPTER 7

# HETEROGENEOUS NETWORK EMBEDDING LEARNING

## 7.1 Overview

In this chapter, we study the problem of representation learning in heterogeneous networks. The unique challenges come from the existence of multiple types of nodes and links, which limit the feasibility of the conventional network embedding techniques. We develop two novel scalable representation learning models, namely *metapath2vec* and *metapath2vec++*. The *metapath2vec* model formalizes meta-path based random walks to construct the heterogeneous neighborhood of a node and then leverages a heterogeneous skip-gram model to perform node embeddings. The *metapath2vec++* model further enables the simultaneous modeling of structural and semantic correlations in heterogeneous networks. Extensive experiments show that *metapath2vec* and *metapath2vec++* are able to not only outperform state-of-the-art embedding models in various heterogeneous network mining tasks, such as node classification, clustering, and similarity search, but also discern the structural and semantic correlations between diverse network objects.

This chapter is largely extracted from a pre-print manuscript (see DCS17 in the NS-CTA publication database). It is a joint work with Nitesh V. Chawla and Ananthram Swami.

## 7.2 Introduction

Neural network based learning models can represent latent embeddings that capture the internal relations from rich, complex data of various modalities, such as image, audio, and language [119]. Social and information networks are rich and complex data encoding the dynamics and modalities of human interactions, and can also be amenable to representation learning using neural networks. In particular, by mapping the way that people choose friends and maintain connections as a "social language," the recent advances in natural language processing (NLP) [15] can be naturally applied to network representation learning. And this has indeed been the case, since the inception of word2vec [145, 146] in NLP. A number of recent research publications have proposed word2vec based network representation learning networks, such as DeepWalk [168], LINE [207], and node2vec [82]. Specifically, instead of handcrafted network feature design, representation learning enables the automatic discovery of useful and meaningful (latent) features from the "raw networks".

However, these work have focused on representation learning for homogeneous networks—representative of singular type of nodes and / or relationships. A number of social and information networks are heterogeneous in nature, involving diversity of node types and relationships between nodes [196]. These heterogeneous networks present unique challenges that cannot be handled by representation learning models that are specifically designed for homogeneous networks. Take a heterogeneous academic network as an example, how do we effectively preserve the concept of "word-context" among multiple types of nodes, e.g., authors, papers, venues, organizations, etc.? Can random walks, such those used in DeepWalk and node2vec, be applied to networks of multiple types of nodes? Can we directly apply homogeneous network oriented embedding architectures (e.g., skip-gram) to heterogeneous networks?

By solving these challenges, the latent heterogeneous network embeddings can be further applied to various network mining tasks, such as node classification [98],

clustering [197, 200], and similarity search [198, 240]. In contrast to conventional meta-path based methods [196], the advantage of latent-space representation learning lies in its ability to model similarities between nodes without connected meta-paths. For example, if authors have never published papers in the same venues—imagine one publishes 10 papers all in NIPS and the other has 10 publications all in ICML, their "APCPA" based PathSim similarity [198] would be zero—an issue that can be naturally conquered by network representation learning.

**Contributions.** We formalize the heterogeneous network representation learning problem, where the objective is to simultaneously learn the low-dimensional and latent embeddings for multiple types of nodes. We present the *metapath2vec* and its extension *metapath2vec++* frameworks. The goal of *metapath2vec* is to maximize the likelihood of preserving both the structures and semantics of a given heterogeneous network. In *metapath2vec*, we first propose meta-path [196] based random walks in heterogeneous networks to generate heterogeneous neighborhoods with network semantics for various types of nodes. Second, we extend the skip-gram model [146] to facilitate the modeling of geographically and semantically close nodes. Finally, we develop a heterogeneous negative sampling based method—referred to as *metapath2vec++*, enabling the accurate and efficient prediction of a node's heterogeneous neighborhood.

The proposed *metapath2vec* and *metapath2vec++* are different from conventional network embedding models, which focus on homogeneous networks [82, 168, 207]. Specifically, they suffer from the identical treatment for different types of nodes and relations, leading to the production of indistinguishable representations for heterogeneous nodes, as also evident through our evaluation. Further, the *metapath2vec* and *metapath2vec++* models also differ from the Predictive Text Embedding (PTE) model [206] in several ways. First, PTE is a semi-supervised learning model that incorporates label information for text data. Second, the heterogeneity in PTE comes

from the text network wherein a link connects two words, a word and its document, and a word and its label. Essentially, the raw input of PTE is words and its output is the embedding of each word, rather than multiple types of objects.

As a highlight, we summarize the differences of these methods in Table 7.1, which lists their input to learning algorithms, as well as the top-five similarity search results in the DBIS network for the same two queries used in [198] (see Section 7.5 for detail). By modeling the heterogeneous neighborhood and further leveraging the heterogeneous negative sampling technique, *metapath2vec++* is able to achieve the best top-five similar results for both types of queries. Figure 7.1 shows the visualization of the 2D projections of the learned embeddings for 16 CS conferences and corresponding high-profile researchers in each field. Remarkably, we find that *metapath2vec++* is capable of automatically organizing these two types of nodes and implicitly learning the internal relationships between them, suggested by the similar directions and distances of the arrows connecting each pair of them, such as J. Dean $\rightarrow$ OSDI and C. D. Manning $\rightarrow$ ACL, and *metapath2vec* is able to group each pair of one author and one conference closely, such as R. E. Tarjan and FOCS. All these properties are not discoverable from conventional network embedding models.

To summarize, our work makes the following contributions:

- Formalizes the problem of heterogeneous network representation learning and identifies its unique challenges resulting from network heterogeneity.
- Develops effective and efficient network embedding frameworks, *metapath2vec* & *metapath2vec++*, for preserving both structural and semantic correlations of heterogeneous networks.
- Through extensive experiments, demonstrates the efficacy and scalability of the presented methods in various heterogeneous network mining tasks, such as node classification (achieving 35–319% improvement over benchmarks) and node clustering (achieving 13–16% gain over baselines).
- Demonstrates the automatic discovery of internal semantic relationships between different types of nodes in heterogeneous networks by *metapath2vec* & *metapath2vec++*, not discoverable by existing work.

TABLE 7.1

CASE STUDY OF SIMILARITY SEARCH IN THE HETEROGENEOUS DBIS DATA USED IN [198]

| Method | PathSim [198] | | DeepWalk/node2vec [82, 168] | | LINE (1st+2nd) [207] | | PTE [206] | | *metapath2vec++* | |
|---|---|---|---|---|---|---|---|---|---|---|
| Input | meta-paths | | heter. rw paths | | heter. edges | | heter. edges | | prob. meta-paths | |
| Query | PKDD | C. Faloutsos | PKDD | C. Faloutsos | PKDD | C. Faloutsos | PKDD | C. Faloutsos | PKDD | C. Faloutsos |
| 1 | ICDM | J. Han | R. S. | J. Pan | W. K. | C. Aggarwal | KDD | C. Aggarwal | KDD | R. Agrawal |
| 2 | SDM | R. Agrawal | M. N. | H. Tong | S. A. | P. Yu | ICDM | P. Yu | PAKDD | J. Han |
| 3 | PAKDD | J. Pei | R. P. | H. Yang | A. B. | D. Gunopulos | SDM | Y. Tao | ICDM | J. Pei |
| 4 | KDD | C. Aggarwal | G. G. | R. Filho | M. S. | N. Koudas | DMKD | N. Koudas | DMKD | C. Aggarwal |
| 5 | DMKD | H. Jagadish | F. J. | R. Chan | S. A. | M. Vlachos | PAKDD | R. Rastogi | SDM | P. Yu |

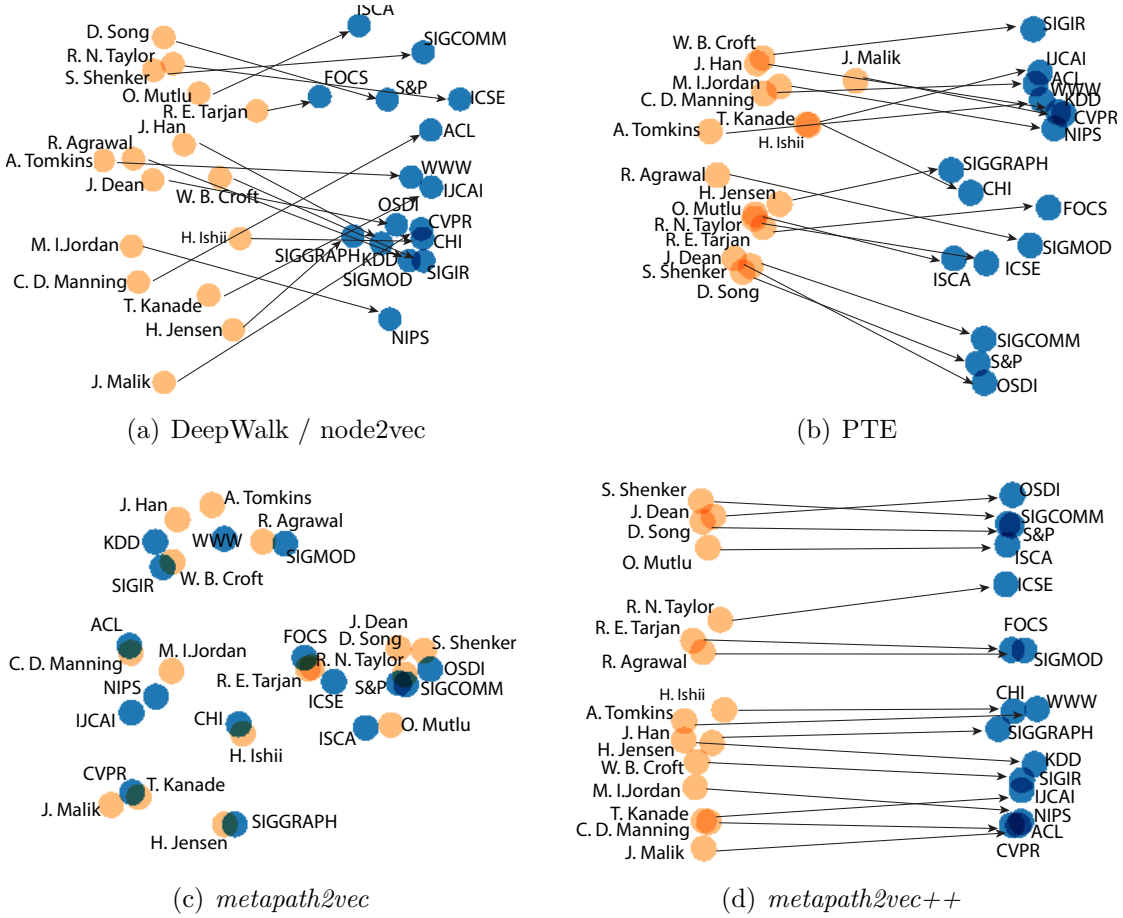



Figure 7.1. 2D PCA projections of the 128-d embeddings of 16 top CS conferences and corresponding high-profile authors learned by DeepWalk / node2vc, PTE, *metapath2vec*, and *metapath2vec++*.

## 7.3 Problem Definition

We formalize the representation learning problem in heterogeneous networks. In specific, we define heterogeneous networks and present the learning problem with its inputs and outputs.

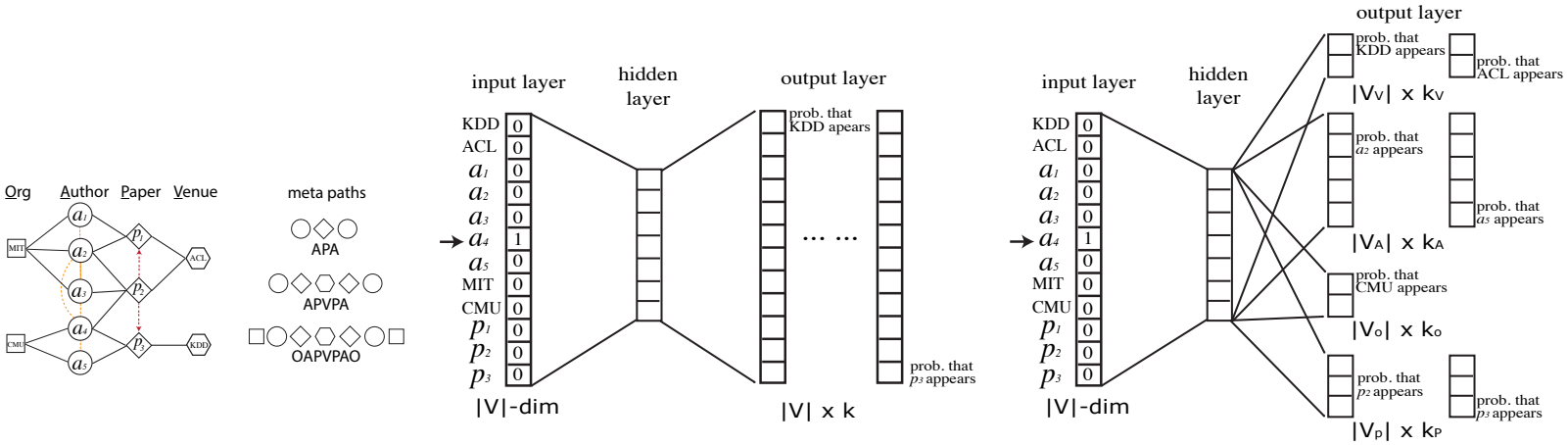
**Definition 6** ***A Heterogeneous Network*** *is defined as a graph* $G = (V, E, T)$ *in which each node* $v$ *and each link* $e$ *are associated with their mapping functions* $\phi(v) : V \to T_V$ *and* $\varphi(e) : E \to T_E$*, respectively.* $T_V$ *and* $T_E$ *denote the sets of object and relation types, where* $|T_V| + |T_E| > 2$*.*

For example, one can represent the academic network in Figure 7.2(a) with authors (A), papers (P), venues (V), organizations (O) as nodes, wherein edges indicate the coauthor (A–A), publish (A–P, P–V), affiliation (O–A) relationships. By considering a heterogeneous network as input, we formalize the problem of heterogeneous network representation learning as follows.

**Problem 5** ***Heterogeneous Network Representation Learning:*** *Given a heterogeneous network $G$, the task is to learn the $d$-dimensional latent representations $\mathbf{X} \in \mathbb{R}^{|V| \times d}$, $d \ll |V|$ that are able to capture the structural and semantic relations among them.*

The output of the problem is the low-dimensional matrix $\mathbf{X}$, with the $v^{th}$ row—a $d$-dimensional vector $X_v$—corresponding to the representation of node $v$. Notice that, although there are different types of nodes in $V$, their representations are mapped into the same latent space. The learned node representations can benefit various heterogeneous network mining tasks. For example, the embedding vector of each node can be used as the feature input of node classification, clustering, and similarity search tasks.

The main challenge of this problem comes from the network heterogeneity, wherein it is difficult to directly apply homogeneous language and network embedding methods. The premise of network embedding models is to preserve the proximity between a node and its neighborhood (context) [82, 168, 207]. In a heterogeneous environment, how do we define and model this 'node–neighborhood' concept? Furthermore, how do we optimize the embedding models that effectively maintain the structures and semantics of multiple types of nodes and relations?




(a) A toy network (b) Skip-gram in *metapath2vec* (c) Skip-gram in *metapath2vec++*

Figure 7.2. An illustrative example of a heterogeneous network and skip-gram architectures of *metapath2vec* and *metapath2vec++* for embedding this network. (a). Yellow dot lines denote coauthor relationships and red dot lines denote citation relationships. (b) The skip-gram architecture used in *metapath2vec* when predicting for $a_4$, which is the same with the one in node2vec if node types are ignored. $|V|$=12 denotes the number of nodes in the heterogeneous academic network in (a) and $a_4$'s neighborhood is set to include CMU, $a_2$, $a_3$, $a_5$, $p_2$, $p_3$, ACL, & KDD, making $k = 8$. (c) The heterogeneous skip-gram used in *metapath2vec++*. Instead of one set of multinomial distributions for all types of neighborhood nodes in the output layer, it specifies one set of multinomial distributions for each type of nodes in $a_4$'s neighborhood. $V_t$ denotes one specific $t$-type nodes and $V = V_V \cup V_A \cup V_O \cup V_P$. $k_t$ specifies the size of a particular type of one's neighborhood and $k = k_V + k_A + k_O + k_P$.

## 7.4 The *metapath2vec* Framework

We present a general framework, *metapath2vec*, which is capable of learning desirable node representations in heterogeneous networks. The objective of *metapath2vec* is to maximize the network probability in consideration of multiple types of nodes and edges.

### 7.4.1 Skip-Gram in Homogeneous Network Embedding

We, first, briefly introduce the word2vec model and its application to homogeneous network embedding tasks. Given a text corpus, Mikolov et al. proposed *word2vec* to learn the distributed representations of words in a corpus [145, 146]. Inspired by it, DeepWalk [168] and node2vec [82] aim to map the word-context concept in a text corpus into a network. Both methods leverage random walks to achieve this and utilize the skip-gram model to learn the representation of a node that facilitates the prediction of its structural context—local neighborhoods—in a homogeneous network. Usually, given a network $G = (V, E)$, the objective is to maximize the network probability in terms of local structures, that is:

$$\arg\max_{\theta} \prod_{v \in V} \prod_{c \in N(v)} p(c|v;\theta) \tag{7.1}$$

where $N(v)$ is the neighborhood of node $v$ in the network $G$, which can be defined in different ways such as $v$'s one-hop neighbors, and $p(c|v;\theta)$ defines the conditional probability of having a context node $c$ given a node $v$.

### 7.4.2 Heterogeneous Network Embedding: *metapath2vec*

To model the heterogeneous neighborhood of a node, *metapath2vec* introduces the heterogeneous skip-gram model. To incorporate the heterogeneous network structures into skip-gram, we propose meta-path based random walks in heterogeneous

networks.

**Heterogeneous Skip-Gram.** In *metapath2vec*, we enable skip-gram to learn effective node representations for a heterogeneous network $G = (V, E, T)$ with $|T_V| > 1$ by maximizing the probability of having the heterogeneous context $N_t(v), t \in T_V$ give a node $v$:

$$\arg\max_{\theta} \sum_{v \in V} \sum_{t \in T_V} \sum_{c_t \in N_t(v)} \log p(c_t|v;\theta) \tag{7.2}$$

where $N_t(v)$ denotes $v$'s neighborhood with the $t^{th}$ type of nodes and $p(c_t|v;\theta)$ is commonly defined as a softmax function [15, 76, 146, 176], that is: $p(c_t|v;\theta) = \frac{e^{X_{c_t} \cdot X_v}}{\sum_{u \in V} e^{X_u \cdot X_v}}$, where $X_v$ is the $v^{th}$ row of $\mathbf{X}$, representing the embedding vector for node $v$. For illustration, consider the academic network in Figure 7.2(a), the neighborhood of one author node $a_4$ can be structurally close to other authors (e.g., $a_2$, $a_3$ & $a_4$), venues (e.g., ACL & KDD), organizations (CMU & MIT), as well as papers (e.g., $p_2$ & $p_3$).

To achieve efficient optimization, Mikolov et al. introduced negative sampling [146], in which a relatively small set of words (nodes) are sampled from the corpus (network) for the construction of softmax. We leverage the same technique for *metapath2vec*. Given a negative sample size $M$, Eq. 7.2 is updated as follows: $\log \sigma(X_{c_t} \cdot X_v) + \sum_{m=1}^{M} \mathbb{E}_{u^m \sim P(u)}[\log \sigma(-X_{u^m} \cdot X_v)]$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ and $P(u)$ is the pre-defined distribution from which a negative node $u^m$ is drew from for $M$ times. *metapath2vec* builds the the node frequency distribution by viewing different types of nodes homogeneously and draw (negative) nodes regardless of their types.

**Meta-Path Based Random Walks.** How to effectively transform the structure of a network into skip-gram? In DeepWalk [168] and node2vec [82], this is achieved by incorporating the node paths traversed by random walkers over a network into the neighborhood function.

Naturally, we can put *random walkers in a heterogeneous network* to generate paths of multiple types of nodes. At step $i$, the transition probability $p(v^{i+1}|v^i)$ is denoted as the normalized probability distributed over the neighbors of $v^i$ by ignoring their node types. The generated paths can be then used as the input of node2vec and DeepWalk. *However, Sun et al. demonstrated that heterogeneous random walks are biased to highly visible types of nodes—those with a dominant number of paths—and concentrated nodes—those with a governing percentage of paths pointing to a small set of nodes [198].*

In light of these issues, we design meta-path based random walks to generate paths that are able to capture both the semantic and structural correlations between different types of nodes, facilitating the transformation of heterogeneous network structures into *metapath2vec*'s skip-gram.

Formally, a meta-path scheme $\mathcal{P}$ is defined as a path that is denoted in the form of $V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \cdots V_t \xrightarrow{R_t} V_{t+1} \cdots \xrightarrow{R_{l-1}} V_l$, wherein $R = R_1 \circ R_2 \circ \cdots \circ R_{l-1}$ defines the composite relations between node types $V_1$ and $V_l$ [196]. Take Figure 7.2(a) as an example, a meta-path "APA" represents the coauthor relationships on a paper (P) between two authors (A), and "APVPA" represents two authors (A) publish papers (P) in the same venue (V). Previous work has shown that many data mining tasks in heterogeneous information networks can benefit from the modeling of meta paths [196].

Here we show how to use meta-paths to guide heterogeneous random walkers. Given a heterogeneous network $G = (V, E, T)$ and a meta-path scheme $\mathcal{P}$: $V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \cdots V_t \xrightarrow{R_t} V_{t+1} \cdots \xrightarrow{R_{l-1}} V_l$, the transition probability at step $i$ is defined as

follows:

$$p(v^{i+1}|v_t^i, \mathcal{P}) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1}, v_t^i) \in E,\ \phi(v^{i+1}) = t{+}1 \\ 0 & (v^{i+1}, v_t^i) \in E,\ \phi(v^{i+1}) \neq t{+}1 \\ 0 & (v^{i+1}, v_t^i) \notin E \end{cases} \tag{7.3}$$

where $v_t^i \in V_t$ and $N_{t+1}(v_t^i)$ denotes the $V_{t+1}$ type of neighborhood of node $v_t^i$. In other words, $v^{i+1} \in V_{t+1}$, that is, the flow of the walker is conditioned on the pre-defined meta-path $\mathcal{P}$. In addition, meta-paths are commonly used in a symmetric way, that is, its first node type $V_1$ is the same with the last one $V_l$ [196–198], facilitating its recursive guidance for random walkers, i.e.,

$$p(v^{i+1}|v_t^i) = p(v^{i+1}|v_1^i),\ \text{if } t = l \tag{7.4}$$

The meta-path based random walk strategy ensures that the semantic relationships between different types of nodes can be properly incorporated into skip-gram. For example, in a traditional random walk procedure, in Figure 7.2(a), the next step of a walker on node $a_4$ transitioned from node CMU can be all types of nodes surrounding it—$a_2$, $a_3$, $a_5$, $p_2$, $p_3$, and CMU. However, under the meta-path scheme 'OAPVPAO', for example, the walker is biased towards paper nodes (P) given its previous step on an organization node CMU (O), following the semantics of this path.

### 7.4.3 The *metapath2vec++* Model

*metapath2vec* distinguishes the context nodes of node $v$ conditioned on their types when constructing its neighborhood function $N_t(v)$ in Eq. 7.2. However, it ignores the node type information in softmax. In other words, in order to infer the specific type of context $c_t$ in $N_t(v)$ given a node $v$, *metapath2vec* actually encourages all types

of negative samples, including nodes of the same type $t$ as well as the other types in the heterogeneous network.

**Heterogeneous negative sampling.** We further propose the *metapath2vec++* framework, in which the softmax function is normalized with respect to the node type of the context $c_t$. Specifically, $p(c_t|v;\theta)$ is adjusted to the specific node type $t$, that is,

$$p(c_t|v;\theta) = \frac{e^{X_{c_t}\cdot X_v}}{\sum_{u_t\in V_t} e^{X_{u_t}\cdot X_v}} \tag{7.5}$$

where $V_t$ is the node set of type $t$ in the network. In doing so, *metapath2vec++* specifies one set of multinomial distributions for each type of neighborhood in the output layer of the skip-gram model. Recall that in *metapath2vec* and node2vec / DeepWalk, the dimension of the output multinomial distributions is equal to the number of nodes in the network. However, in *metapath2vec++*'s skip-gram, the multinomial distribution dimension for type $t$ nodes is determined by the number of $t$-type nodes. A clear illustration can be seen in Figure 7.2(c). For example, given the target node $a_4$ in the input layer, *metapath2vec++* outputs four sets of multinomial distributions, each corresponding to one type of neighbors—venues $V$, authors $A$, organizations $O$, and papers $P$.

Inspired by PTE [206], the sampling distribution is also specified by the node type of the neighbor $c_t$ that is targeted to predict, i.e., $P_t(\cdot)$. Therefore, we have the following objective:

$$\mathcal{O}(\mathbf{X}) = \log\sigma(X_{c_t}\cdot X_v) + \sum_{m=1}^{M} \mathbb{E}_{u_t^m\sim P_t(u_t)}[\log\sigma(-X_{u_t^m}\cdot X_v)] \tag{7.6}$$

whose gradients are derived as follows:

$$\frac{\partial \mathcal{O}(\mathbf{X})}{\partial X_{u_t^m}} = (\sigma(X_{u_t^m} \cdot X_v - \mathbb{I}_{c_t}[u_t^m]))X_v \tag{7.7}$$
$$\frac{\partial \mathcal{O}(\mathbf{X})}{\partial X_v} = \sum_{m=0}^{M} (\sigma(X_{u_t^m} \cdot X_v - \mathbb{I}_{c_t}[u_t^m]))X_{u_t^m}$$

where $\mathbb{I}_{c_t}[u_t^m]$ is an indicator function to indicate whether $u_t^m$ is the neighborhood context node $c_t$ and when $m = 0$, $u_t^0 = c_t$. The model is optimized by using stochastic gradient descent algorithm. The pseudo code of *metapath2vec++* is listed in Algorithm 2.

## 7.5 Experiments

In this section, we demonstrate the efficacy and efficiency of the presented *metapath2vec* and *metapath2vec++* frameworks for heterogeneous network representation learning.

**Data.** We use two heterogeneous networks, including the AMiner Computer Science (CS) dataset [203] and the Database and Information Systems (DBIS) dataset [198]. Both datasets are publicly available.

This AMiner CS dataset consists of 9,323,739 computer scientists and 3,194,405 papers from 3,883 computer science venues—both conferences and journals—held until 2016. We construct a heterogeneous collaboration network, in which there are three types of nodes: authors, papers, and venues. The links represent different types of relationships among three sets of nodes—such as collaboration relationships on a paper.

The DBIS dataset was constructed and used by Sun et al. [198]. It covers 464 venues, their top-5000 authors, and corresponding 72,902 publications. We also construct the heterogeneous collaboration networks from DBIS wherein a link may con-

**ALGORITHM 2:** The *metapath2vec++* Algorithm.

**Input:** The heterogeneous information network $G = (V, E, T)$, a meta-path scheme $\mathcal{P}$, #walks per node $w$, walk length $l$, embedding dimension $d$, neighborhood size $k$

**Output:** The latent node embeddings $\mathbf{X} \in \mathbb{R}^{|V| \times d}$

initialize $\mathbf{X}$ ;
**for** $i = 1 \rightarrow w$ **do**
  **for** $v \in V$ **do**
    $MP$ = MetaPathRandomWalk($G$, $\mathcal{P}$, $v$, $l$) ;
    $\mathbf{X}$ = HeterogeneousSkipGram($\mathbf{X}$, $k$, $MP$) ;
  **end**
**end**
return $\mathbf{X}$ ;

**MetaPathRandomWalk**($G$, $\mathcal{P}$, $v$, $l$)
$MP$[1] = $v$ ;
**for** $i = 1 \rightarrow l-1$ **do**
  draw u according to Eq. 7.3 ;
  $MP$[i+1] = $u$ ;
**end**
return $MP$ ;

**HeterogeneousSkipGram**($\mathbf{X}$, $k$, $MP$)
**for** $i = 1 \rightarrow l$ **do**
  $v = MP$[i] ;
  **for** $j = max(0, i-k) \rightarrow min(i+k, l)$ & $j \neq i$ **do**
    $c_t = MP$[j] ;
    $X^{new} = X^{old} - \eta \cdot \frac{\partial \mathcal{O}(\mathbf{X})}{\partial X}$ (Eq. 7.7) ;
  **end**
**end**

nect two authors, one author and one paper, as well as one paper and one venue.

### 7.5.1 Experimental Setup

We compare—*metapath2vec* and *metapath2vec++*—with several recent network representation learning methods:

- DeepWalk [168] / node2vec [82]: With the same random walk path input ($p$=1 & $q$=1 in node2vec), we find that the choice between hierarchical softmax (DeepWalk) and negative sampling (node2vec) techniques does not yield

significant differences. Therefore we use $p$=1 and $q$=1 [82] in node2vec for comparison.

- LINE [207]: We use the advanced version of LINE by considering both the 1st and 2nd order of node proximity;
- PTE [206]: We construct three bipartite heterogeneous networks—author–author, author–venue, venue–venue—and restrain it as an unsupervised embedding method;
- Spectral Clustering [209] / Graph Factorization [4]: With the same treatment to these methods in the node2vec work [82], we make them excluded for comparison, as previous studies have demonstrated the outperformance of DeepWalk and LINE to them.

For all embedding methods, we use the same parameters listed below. In addition, we also vary each of them and fix the others for examining the parameter sensitivity of the proposed methods.

- The number of walks per node $w$: 1000;
- The walk length $l$: 100;
- The vector dimension $d$: 128 (LINE: 128 for each order);
- The neighborhood size $k$: 7;
- The size of negative samples: 5.

For *metapath2vec* and *metapath2vec++*, we also need to specify the meta-path scheme to guide random walks. We surveyed most of the meta-path based work and found that the most commonly and effectively used meta-path schemes in heterogeneous academic networks are "APA" and "APVPA" [196, 198, 200]. Notice that "APA" denotes the coauthor semantic, that is, the traditional (homogeneous) collaboration links / relationships. "APVPA" represents the heterogeneous semantic of 'authors publish papers at the same venues'. Our empirical results also show that this simple meta-path scheme "APVPA" can lead to node embeddings that can be generalized to diverse heterogeneous academic mining tasks, suggesting its applicability to potential applications for academic search services.

We evaluate the quality of the latent representations learned by different methods over three classical heterogeneous network mining tasks, including multi-class node classification [98], node clustering [200], and similarity search [198]. In addition, we also use the embedding projector in TensorFlow [1] to visualize the node embeddings learned from the heterogeneous academic networks.

### 7.5.2 Multi-Class Classification

For the classification task, we use third-party labels to determine the class of each node. First, we match the following eight categories of venues in Google Scholar with those in AMiner data: Computational Linguistics, Computer Graphics, Computer Networks & Wireless Communication, Computer Vision & Pattern Recognition, Computing Systems, Databases & Information Systems, Human Computer Interaction, and Theoretical Computer Science. Among all of the 160 venues (20 per category $\times$ 8 categories), 133 of them are successfully matched and labeled correspondingly (Most of unmatched venues are pre-print venues, such as arXiv). Second, for each author who published in these 133 venues, his / her label is assigned to the category with the majority of his / her publications, and a tie is resolved by random selection among the possible categories; 246,678 authors are labeled with research category.

Note that the node representations are learned from the full dataset. The embeddings of above labeled nodes are then used as the input to a logistic regression classifier. In the classification experiments, we vary the size of the training set from 5% to 50% and the remaining nodes for testing. We repeat each prediction experiment ten times and report the average performance in terms of both Macro-F1 and Micro-F1 scores.

TABLE 7.2

MULTI-CLASS VENUE CLASSIFICATION RESULTS (F1) IN AMINER DATA

| Metric | Method | 5% | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Macro-F1 | node2vec | 0.0723 | 0.1396 | 0.1905 | 0.2795 | 0.3427 | 0.3911 | 0.4424 | 0.4774 | 0.4955 | 0.4457 |
| | LINE(1st+2nd) | 0.2245 | 0.4629 | 0.7011 | 0.8473 | 0.8953 | 0.9203 | 0.9308 | 0.9466 | 0.9410 | 0.9466 |
| | PTE | 0.1702 | 0.3388 | 0.6535 | 0.8304 | 0.8936 | 0.9210 | 0.9352 | 0.9505 | 0.9525 | 0.9489 |
| | *metapath2vec* | 0.3033 | 0.5247 | 0.8033 | 0.8971 | 0.9406 | 0.9532 | 0.9529 | 0.9701 | 0.9683 | 0.9670 |
| | *metapath2vec++* | 0.3090 | 0.5444 | 0.8049 | 0.8995 | 0.9468 | 0.9580 | 0.9561 | 0.9675 | 0.9533 | 0.9503 |
| Micro-F1 | node2vec | 0.1701 | 0.2142 | 0.2486 | 0.3266 | 0.3788 | 0.4090 | 0.4630 | 0.4975 | 0.5259 | 0.5286 |
| | LINE(1st+2nd) | 0.3000 | 0.5167 | 0.7159 | 0.8457 | 0.8950 | 0.9209 | 0.9333 | 0.9500 | 0.9556 | 0.9571 |
| | PTE | 0.2512 | 0.4267 | 0.6879 | 0.8372 | 0.8950 | 0.9239 | 0.9352 | 0.9550 | 0.9667 | 0.9571 |
| | *metapath2vec* | 0.4173 | 0.5975 | 0.8327 | 0.9011 | 0.9400 | 0.9522 | 0.9537 | 0.9725 | 0.9815 | 0.9857 |
| | *metapath2vec++* | 0.4331 | 0.6192 | 0.8336 | 0.9032 | 0.9463 | 0.9582 | 0.9574 | 0.9700 | 0.9741 | 0.9786 |

TABLE 7.3

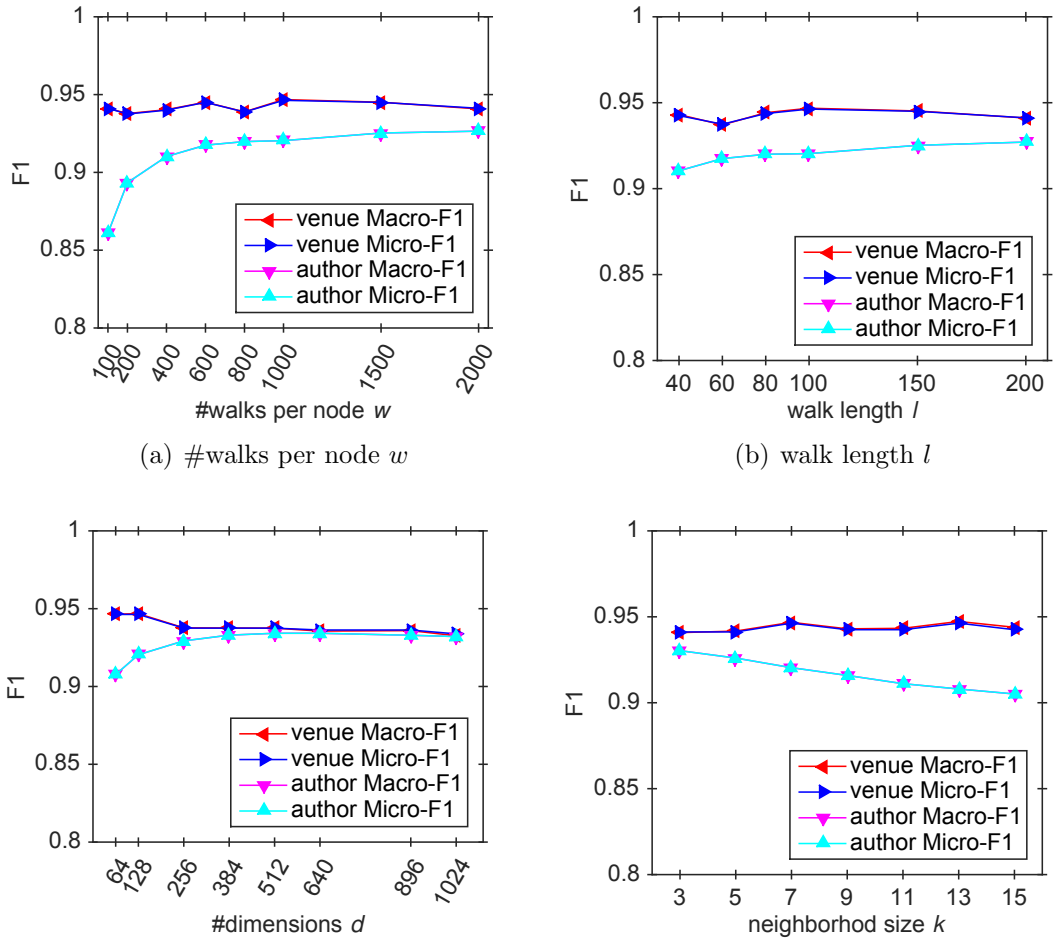
MULTI-CLASS AUTHOR CLASSIFICATION RESULTS (F1) IN AMINER DATA

| Metric | Method | 5% | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Macro-F1 | node2vec | 0.7153 | 0.7222 | 0.7256 | 0.7270 | 0.7273 | 0.7274 | 0.7273 | 0.7271 | 0.7275 | 0.7275 |
| | LINE(1st+2nd) | 0.8849 | 0.8886 | 0.8911 | 0.8921 | 0.8926 | 0.8929 | 0.8934 | 0.8936 | 0.8938 | 0.8934 |
| | PTE | 0.8898 | 0.8940 | 0.897 | 0.8982 | 0.8987 | 0.8990 | 0.8997 | 0.8999 | 0.9002 | 0.9005 |
| | *metapath2vec* | 0.9216 | 0.9262 | 0.9292 | 0.9303 | 0.9309 | 0.9314 | 0.9315 | 0.9316 | 0.9319 | 0.9320 |
| | *metapath2vec++* | 0.9107 | 0.9156 | 0.9186 | 0.9199 | 0.9204 | 0.9207 | 0.9207 | 0.9208 | 0.9211 | 0.9212 |
| Micro-F1 | node2vec | 0.7312 | 0.7372 | 0.7402 | 0.7414 | 0.7418 | 0.7420 | 0.7419 | 0.7420 | 0.7425 | 0.7425 |
| | LINE(1st+2nd) | 0.8936 | 0.8969 | 0.8993 | 0.9002 | 0.9007 | 0.9010 | 0.9015 | 0.9016 | 0.9018 | 0.9017 |
| | PTE | 0.8986 | 0.9023 | 0.9051 | 0.9061 | 0.9066 | 0.9068 | 0.9075 | 0.9077 | 0.9079 | 0.9082 |
| | *metapath2vec* | 0.9279 | 0.9319 | 0.9346 | 0.9356 | 0.9361 | 0.9365 | 0.9365 | 0.9365 | 0.9367 | 0.9369 |
| | *metapath2vec++* | 0.9173 | 0.9217 | 0.9243 | 0.9254 | 0.9259 | 0.9261 | 0.9261 | 0.9262 | 0.9264 | 0.9266 |

**Results.** Tables 7.2 and 7.3 list the eight-class classification results for venues and authors, respectively. Overall, the proposed *metapath2vec* and *metapath2vec++* models consistently and significantly outperform all baselines in terms of both metrics. When predicting for the venue category, the advantage of both *metapath2vec* and *metapath2vec++* are particular strong given a small size of training data. Given 5% of nodes as training data, for example, *metapath2vec* and *metapath2vec++* achieve 35–319% improvement in terms of Macro-F1 and 39–145% gain in terms of Micro-F1 over DeepWalk / node2vec, LINE, and PTE. When predicting for authors' categories, the performance of each method is relatively stable when varying the train-test split. The constant gain achieved by the proposed methods is around 2–3% over LINE and PTE, and ∼20% over DeepWalk / node2vec.

In summary, *metapath2vec* and *metapath2vec++* learn significantly better heterogeneous node embeddings than current state-of-the-art methods, as measured by multi-class classification performance. The advantage of the proposed methods lies in their proper consideration and accommodation of the network heterogeneity challenge—the existence of multiple types of nodes and relations.

**Parameter sensitivity.** In skip-gram based representation learning models, there exist several common parameters (see Section 7.5.1). We conduct a sensitivity analysis of *metapath2vec++* to these parameters. Figure 7.3 shows the classification results as a function of one chosen parameter when the others are controlled for. In general, we find that in Figures 7.3(a) and 7.3(b) the number of walks $w$ rooting from each node and the length $l$ of each walk are positive to the author classification performance, while they are surprisingly inconsequential for inferring venue nodes' categories as measured by Macro-F1 and Micro-F1 scores. The increase of author classification performance converges as $w$ and $l$ reach around 1000 and 100, respectively. Similarly, Figures 7.3(c) and 7.3(d) suggest that the number of embedding dimensions $d$ and neighborhood size $k$ are again of relatively little relevance to the




(a) #walks per node $w$ (b) walk length $l$

(c) #dimensions $d$ (d) neighborhood size $k$

Figure 7.3. Parameter sensitivity in multi-class node classification. 50% as training data and the remaining as test data.

predictive task for venues, and $k$ on the other hand is positively crucial to determine the class of a venue. However, the descending lines as the increase of $k$ for author classifications imply that a smaller neighborhood size actually produces the best embeddings for separating authors. This finding differs from those in a homogeneous environment [82], wherein the neighborhood size generally shows a positive effect on node classification.

According to the analysis, *metapath2vec++* is not strictly sensitive to these parameters and is able to reach high performance under a cost-effective parameter choice

TABLE 7.4

NODE CLUSTERING RESULTS (NMI) IN THE AMINER DATA

| methods | venue | author |
|---|---|---|
| node2vec | 0.1952 | 0.2941 |
| LINE (1st+2nd) | 0.8967 | 0.6423 |
| PTE | 0.9060 | 0.6483 |
| *metapath2vec* | 0.9274 | 0.7470 |
| *metapath2vec++* | 0.9261 | 0.7354 |

(the smaller, the more efficient). In addition, our results also indicate that those common parameters show different functions for heterogeneous network embedding with those in homogeneous network cases, demonstrating the request of different ideas and solutions for heterogeneous network representation learning.

### 7.5.3 Node Clustering

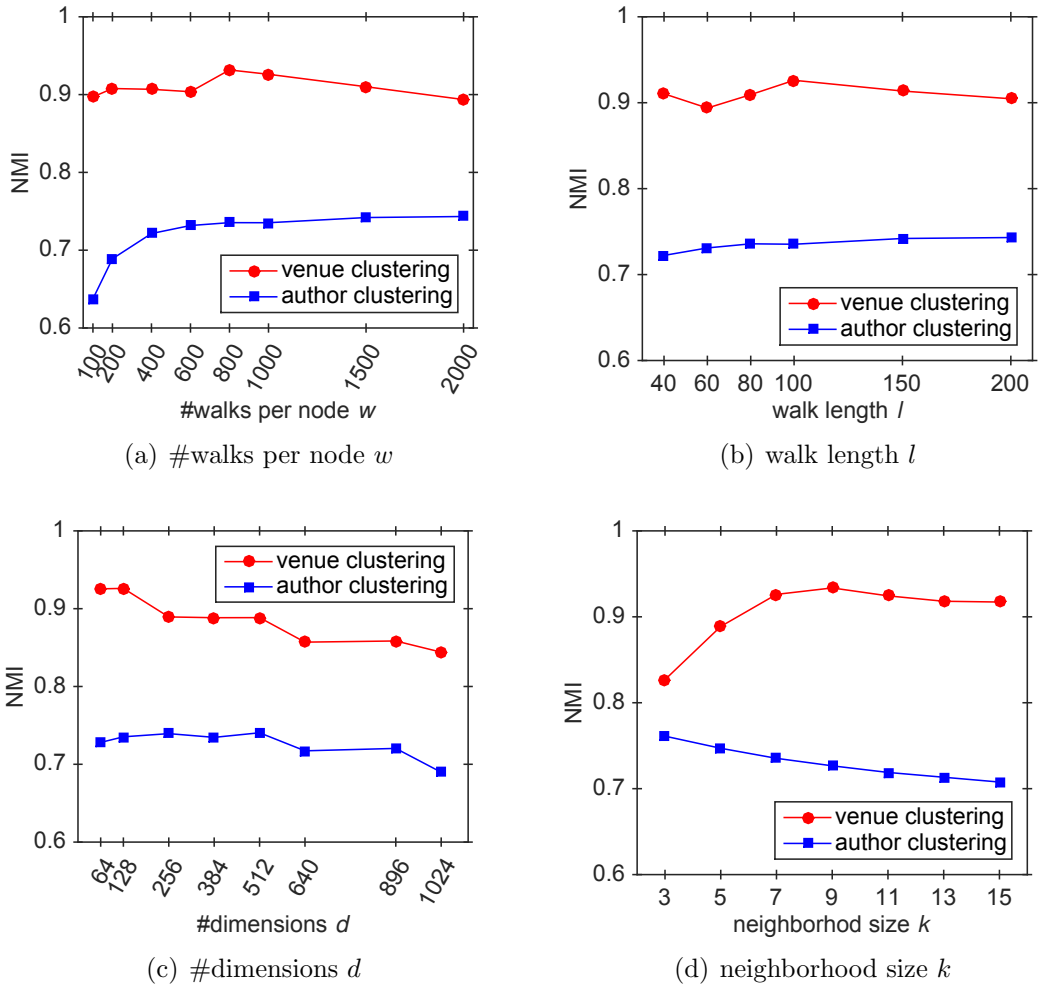
We illustrate how the latent representations learned by embedding methods can help the node clustering task in heterogeneous networks. We employ the same eight-category author and venue nodes used in the classification task above. The learned embeddings by each method is input to a clustering model. Here we leverage the $k$-means algorithm to cluster the data and evaluate the clustering results in terms of normalized mutual information (NMI) [198]. In addition, we also report *metapath2vec++*'s sensitivity with respect to different parameter choices. All clustering experiments are conducted 10 times and the average performance is reported.

**Results.** Table 7.4 shows the node clustering results as measured by NMI in the AMiner CS data. Overall, the table demonstrates that *metapath2vec* and *metap-*

*ath2vec++* outperforms all the comparative methods. When clustering for venues, the task is trivial as evident from the high NMI scores produced by most of the methods: *metapath2vec*, *metapath2vec++*, LINE, and PTE. Nevertheless, the proposed two methods outperform LINE and PTE by 2–3%. The author clustering task is more challenging than the venue case, and the gain obtained by *metapath2vec* and *metapath2vec++* over the best baselines (LINE and PTE) is more significant—around 13–16%.

In summary, *metapath2vec* and *metapath2vec++* generate more appropriate embeddings for different types of nodes in the network than comparison baselines, suggesting their ability to capture and incorporate the underlying structural and semantic relationships between various types of nodes in heterogeneous networks.

**Parameter sensitivity.** Following the same experimental procedure in classification, we study the parameter sensitivity of *metapath2vec++* as measured by the clustering performance. Figure 7.4 shows the clustering performance as a function of each of the four parameters when fixing the other three. From Figures 7.4(a) and 7.4(b), we can observe that the balance between computational cost (a small $w$ and $l$ in $x$-axis) and efficacy (a high NMI in $y$-axis) can be achieved at around $w$ = 800$\sim$1000 and $l$ = 100 for the clustering of both authors and venues. Further, different from the positive effect of increasing $w$ and $l$ on author clustering, $d$ and $k$ are negatively correlated with the author clustering performance, as observed from Figures 7.4(c) and 7.4(d). Similarly, the venue clustering performance also shows an descending trend with an increasing $d$, while on the other hand, we observe a first-increasing and then-decreasing NMI line when $k$ is increased. Both figures together imply that $d$ = 128 and $k$ = 7 are capable of embedding heterogeneous nodes into latent space for promising clustering outcome.




(a) #walks per node $w$ (b) walk length $l$

(c) #dimensions $d$ (d) neighborhood size $k$

Figure 7.4. Parameter sensitivity in clustering.

### 7.5.4 Case Study: Similarity Search

We conduct two case studies to demonstrate the efficacy of our methods. We first select 5 top conferences from the four covered fields in the DBIS data as query nodes, and then select one top conference each from the 16 computer science fields in the AMiner full CS data. We use cosine similarity to determine the distance (similarity) between the query node and the remaining others.

Tables 7.6 and 7.7 list the top ten similar results for querying the 16 leading conferences in corresponding computer science sub-fields in AMiner data. One can

observe that for the query "ACL", for example, *metapath2vec++* returns venues with the same focus—natural language processing, such as EMNLP (1), NAACL (2), Computational Linguistics (3), CoNLL (4), COLING (5), and so on. Similar performance can be also achieved when querying all the other conferences from various domains. More surprisingly, we find that in most cases, the top three results cover venues with similar prestige to the query one, such as STOC to FOCS in theory, OSDI to SOSP in system, HPCA to ISCA in architecture, CCS to S&P in security, CSCW to CHI in human-computer interaction, EMNLP to ACL in NLP, ICML to NIPS in machine learning, WSDM to WWW in Web, AAAI to IJCAI in artificial intelligence, VLDB to SIGMOD in database, etc. Similar results can also be observed in Tables 7.5 and 7.1, which show the similarity search results in the DBIS network.

TABLE 7.5

CASE STUDY OF COMPUTER SCIENCE VENUE SIMILARITY SEARCH IN DBIS DATA

| Rank | KDD | SIGMOD | SIGIR | WWW | WSDM |
|---|---|---|---|---|---|
| 0 | KDD | SIGMOD | SIGIR | WWW | WSDM |
| 1 | SDM | PVLDB | TREC | CIKM | WWW |
| 2 | ICDM | ICDE | CIKM | SIGIR | SIGIR |
| 3 | DMKD | TODS | IPM | KDD | KDD |
| 4 | KDD E | VLDBJ | IRJ | ICDE | AIRWeb |
| 5 | PKDD | PODS | ECIR | TKDE | CIKM |
| 6 | PAKDD | EDBT | TOIS | VLDB | WebDB |
| 7 | TKDE | CIDR | WWW | TOTT | ICDM |
| 8 | CIKM | TKDE | JASIST | SIGMOD | VLDB |
| 9 | ICDE | ICDT | JASIS | WebDB | VLDBJ |
| 10 | TKDD | DE Bull | SIGIRF | WISE | SDM |

TABLE 7.6

CASE STUDY I OF CS VENUE SIMILARITY SEARCH IN AMINER DATA

| Area | Theory | System | Arch | Security | Software | Graphics | Commun. | HCI |
|---|---|---|---|---|---|---|---|---|
| Rank | FOCS | SOSP | ISCA | S&P | ICSE | SIGGRAPH | SIGCOMM | CHI |
| 0 | FOCS | SOSP | ISCA | S&P | ICSE | SIGGRAPH | SIGCOMM | CHI |
| 1 | STOC | TOCS | HPCA | CCS | TOSEM | TOG | CCR | CSCW |
| 2 | SICOMP | OSDI | MICRO | NDSS | FSE | SI3D | HotNets | TOCHI |
| 3 | SODA | HotOS | ASPLOS | USENIX S | ASE | RT | NSDI | UIST |
| 4 | A-R | SIGOPS E | PACT | ACSAC | ISSTA | CGF | CoNEXT | DIS |
| 5 | TALG | ATC | ICS | JCS | E SE | NPAR | IMC | HCI |
| 6 | ICALP | NSDI | HiPEAC | ESORICS | MSR | Vis | TON | MobileHCI |
| 7 | ECCC | OSR | PPOPP | TISS | ESEM | JGT | INFOCOM | INTERACT |
| 8 | TOC | ASPLOS | ICCD | ASIACCS | A SE | VisComp | PAM | GROUP |
| 9 | JAlG | EuroSys | CGO | RAID | ICPC | GI | MobiCom | NordiCHI |
| 10 | ITCS | SIGCOMM | ISLPED | CSFW | WICSA | CG | IPTPS | UbiComp |

TABLE 7.7

CASE STUDY II OF CS VENUE SIMILARITY SEARCH IN AMINER DATA

| Area | NLP | ML | DM | Web | AI | Database | IR | Vision |
|---|---|---|---|---|---|---|---|---|
| Rank | ACL | NIPS | KDD | WWW | IJCAI | SIGMOD | SIGIR | CVPR |
| 0 | ACL | NIPS | KDD | WWW | IJCAI | SIGMOD | SIGIR | CVPR |
| 1 | EMNLP | ICML | SDM | WSDM | AAAI | PVLDB | ECIR | ECCV |
| 2 | NAACL | AISTATS | TKDD | CIKM | AI | ICDE | CIKM | ICCV |
| 3 | CL | JMLR | ICDM | TWEB | JAIR | DE Bull | IRJ | IJCV |
| 4 | CoNLL | NC | DMKD | ICWSM | ECAI | VLDBJ | TREC | ACCV |
| 5 | COLING | MLJ | KDD E | HT | KR | EDBT | SIGIRF | CVIU |
| 6 | IJCNLP | COLT | WSDM | SIGIR | AI Mag | TODS | ICTIR | BMVC |
| 7 | NLE | UAI | CIKM | KDD | ICAPS | CIDR | WSDM | ICPR |
| 8 | ANLP | KDD | PKDD | TIT | CI | SIGMOD R | TOIS | EMMCVPR |
| 9 | LREC | CVPR | ICML | WISE | AIPS | WebDB | IPM | T on IP |
| 10 | EACL | ECML | PAKDD | WebSci | UAI | PODS | AIRS | WACV |

### 7.5.5 Case Study: Visualization

We employ the TensorFlow embedding projector to further visualize the low-dimensional node representations learned by embedding models. First, we project multiple types of nodes—16 top CS conferences and corresponding top-profile authors—into the same space in Figure 7.1. From Figure 7.1(d), we can clearly see that *metapath2vec++* is able to automatically organize these two types of nodes and implicitly learn the internal relationships between them, indicated by the similar directions and distances of the arrows connecting each pair of them, such as J. Dean → OSDI, C. D. Manning → ACL, R. E. Tarjan → FOCS, M. I. Jordan → NIPS, and so on. In addition, these two types of nodes are clearly located in two separate and straight columns. Neither of these two remarkable results can be made by the recent network embedding models in Figures 7.1(a) and 7.1(b).

As to *metapath2vec*, instead of separating the two types of nodes into two columns, it is capable of grouping each pair of one venue and its corresponding author closely, such as R. E. Tarjan and FOCS, H. Jensen and SIGGRAPH, H. Ishli and CHI, R. Agrawal and SIGMOD, etc. Together, both models arrange nodes from similar fields close to each other and dissimilar ones distant from each other, such as the "Core CS" cluster of systems (OSDI), networking (SIGCOMM), security (S&P), and architecture (ISCA), as well as the "Big AI" cluster of data mining (KDD), information retrieval (SIGIR), artificial intelligence (AI), machine learning (NIPS), NLP (ACL), and vision (CVPR). These groupings are also reflected by their corresponding author nodes.

Second, Figure 7.5 visualizes the latent vectors—learned by *metapath2vec++*—of 48 venues used in similarity search of Section 7.5.4, three each from 16 sub-fields. We can see that conferences from the same domain are geographically grouped to each other and each group is well separated from others, further demonstrating the embedding ability of *metapath2vec++*. In addition, the cosine similarity between each pair of venues is presented in Figure 7.6, further demonstrating the embedding

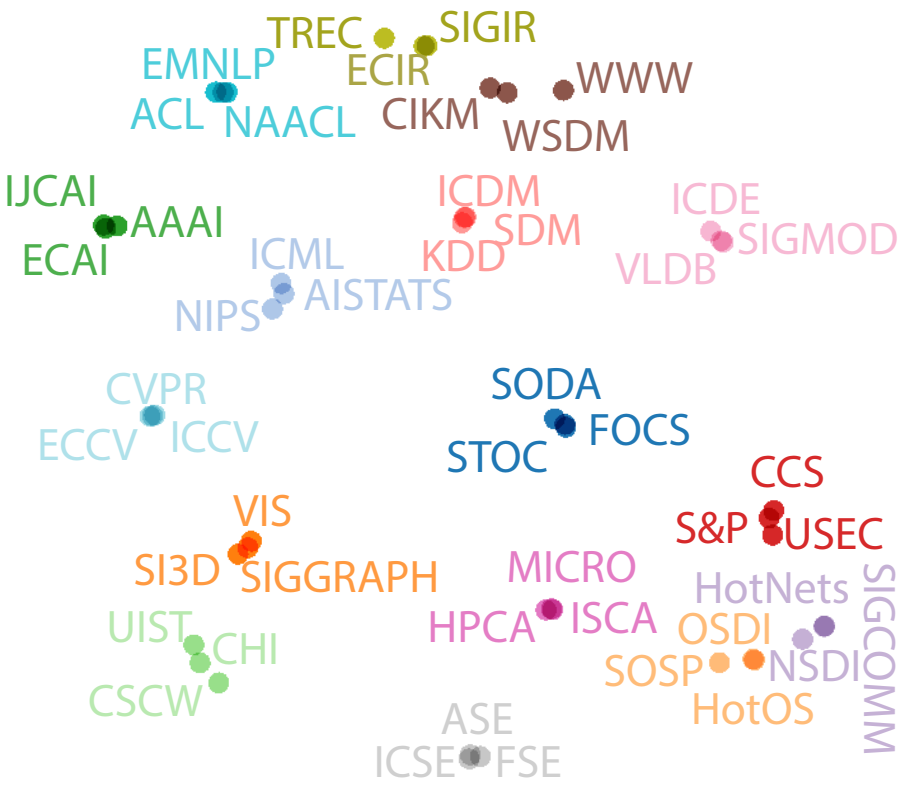



Figure 7.5. 2D t-SNE projections of the 128-d embeddings learned by *metapath2vec++* of 48 CS venues, three each from 16 sub-fields.

ability of *metapath2vec++*. Similar to the observation in Figure 7.1, we can also notice that the heterogeneous embeddings are able to unveil the similarities across different domains, including the "Core CS" sub-field cluster at the bottom right and the "Big AI" sub-field clusters at the top right.

Finally, Figures 7.7 and 7.8 plot the 2D t-SNE projections of eight-category venues and authors used in the node classification and clustering tasks in Sections 7.5.2 and 7.5.3, respectively. When visualizing the eight-category venues (Figure 7.7), both proposed methods and LINE / PTE demonstrate their abilities to separate each category of nodes with others, while DeepWalk and node2vec do not. In general, it is difficult to tell the best separation for venues among LINE, PTE, *metapath2vec*, and *metapath2vec++*. When it comes to author visualization, Figure 7.8 suggests

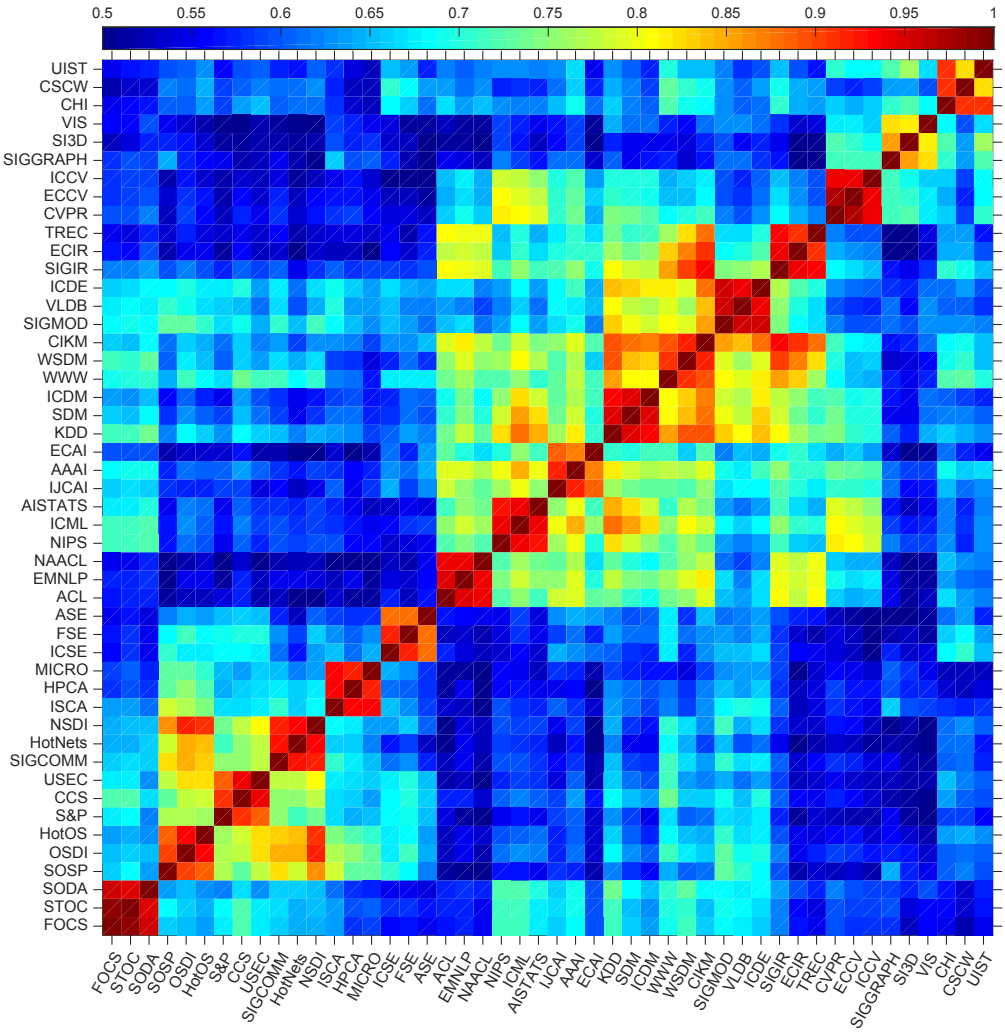
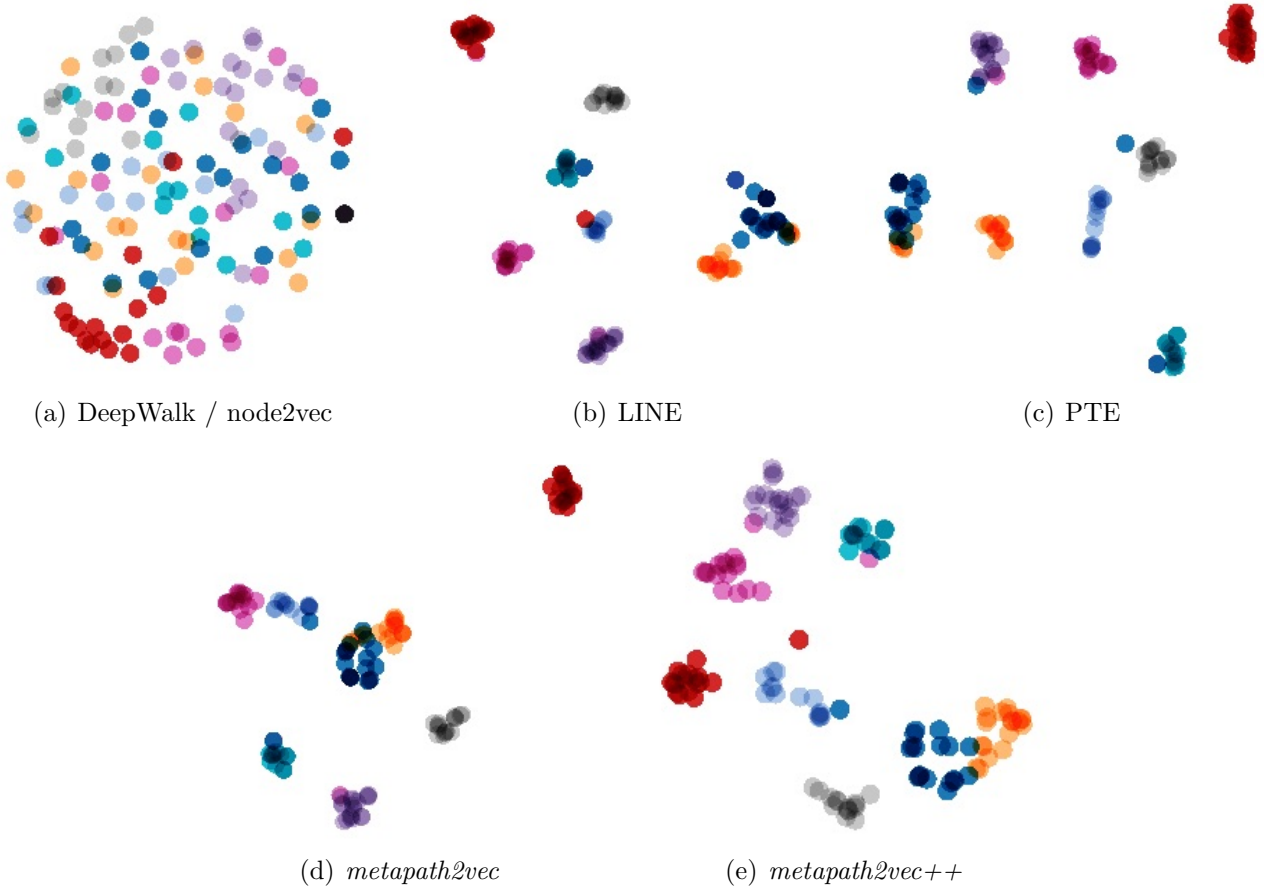



Figure 7.6. Cosine similarity between 48 CS venues, three each from 16 sub-fields.

that the eight clusters provided by *metapath2vec* and *metapath2vec++* are relatively more visible and separable than the other baselines.

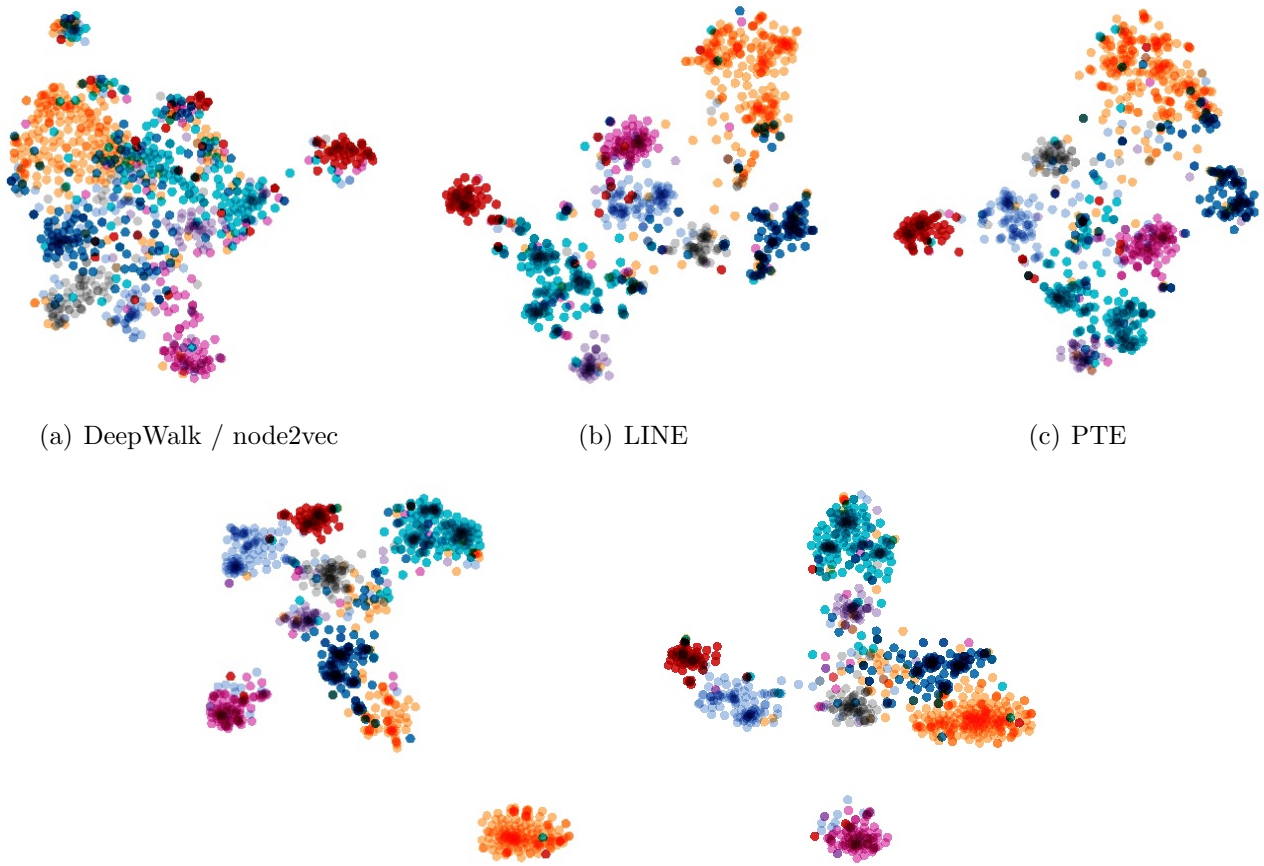
All together, the visualization intuitively demonstrates *metapath2vec++*'s novel capability to discover, model, and capture the underlying structural and semantic relationships between multiple types of nodes in heterogeneous networks.



(a) DeepWalk / node2vec (b) LINE (c) PTE

(d) *metapath2vec* (e) *metapath2vec++*

Figure 7.7. t-SNE visualization of 133 venues in the 8-category data. For all plots, the same parameters—perplexity: 20, learning rate: 1, and #iterations: 2000—are used in TensorFlow online embedding projector.



(a) DeepWalk / node2vec (b) LINE (c) PTE

(d) *metapath2vec* (e) *metapath2vec++*

Figure 7.8. t-SNE visualization of 10,000 randomly sampled authors from the 8-category data. For all plots, the same parameters—perplexity: 20, learning rate: 10, and #iterations: 2000—are used in TensorFlow online embedding projector.

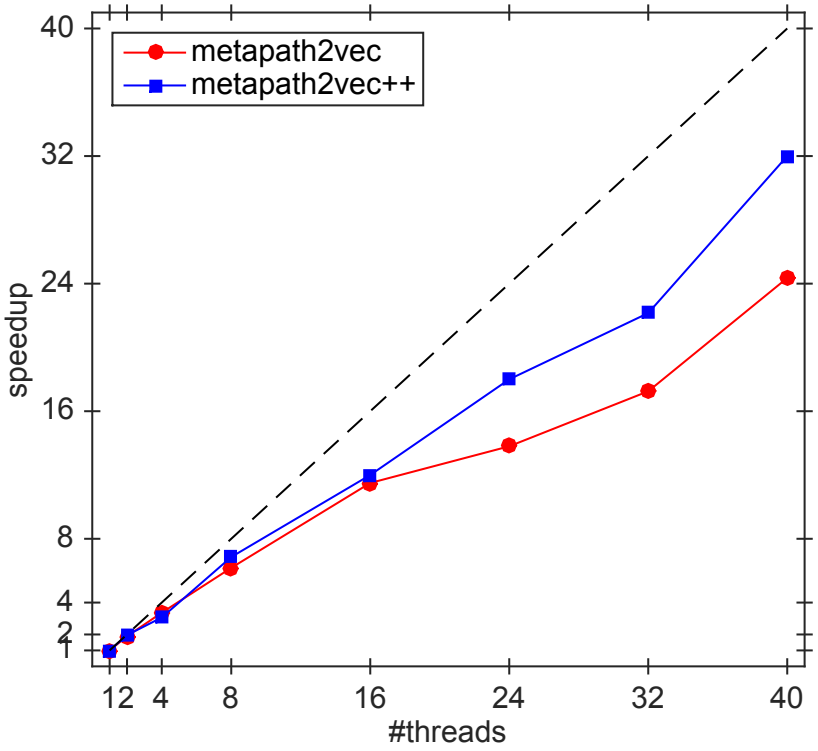



Figure 7.9. Scalability of *metapath2vec* and *metapath2vec++*.

### 7.5.6 Scalability

In the era of big (network) data, it is necessary to demonstrate the scalability of the proposed network embedding models. The *metapath2vec* and *metapath2vec++* methods can be parallelized by using the same mechanism of word2vec and node2vec [82, 146]. All codes are implemented in C and C++ and our experiments are conducted in a computing server with Quad 12 (48) core 2.3 GHz Intel Xeon CPUs E7-4850. We run experiments on the AMiner CS data with the default parameters with different number of threads, i.e., 1, 2, 4, 8, 16, 24, 32, 40, each of them utilizing one CPU core.

Figure 7.9 shows the speedup of *metapath2vec* & *metapath2vec++* over the one-thread running case. Optimal speedup performance is denoted by the dashed $y = x$ line, which represents perfect distribution and execution of computation across

all CPU cores. In general, we find that both methods achieve acceptable sublinear speedups as both lines are close to the optimal line. In specific, they can reach 11–12× speedup with 16 cores and 24–32× speedup with 40 cores used. By using 40 cores, *metapath2vec++*'s learning process costs only 9 minutes for embedding the full AMiner CS network, which is composed of over 9 million authors' 3 million papers published in more than 3800 venues. Overall, the proposed *metapath2vec* and *metapath2vec++* models are efficient and scalable for large-scale heterogeneous networks with millions of nodes.

## 7.6 Related Work

Network representation learning can be traced back to the usage of latent factor models for network analysis and graph mining tasks [91, 235], such as the application of factorization models for recommendation systems [109, 132], node classification [210], relational mining [160], and role discovery [87]. This rich line of research focuses on factorizing the matrix / tensor format (e.g., the adjacency matrix) of a network, generating latent-dimension features for nodes or edges in this network. However, the computational cost of decomposing a large-scale matrix/tensor is usually very expensive, and also suffers from its statistical performance drawback [82], making it neither practical nor effective for addressing tasks in big networks.

With the advent of deep learning techniques, significant effort has been devoted to designing neural network based representation learning models. For example, Mikolov et al. proposed the *word2vec* framework—a two-layer neural network—to learn the distributed representations of words in natural language [145, 146]. Building on word2vec, Perozzi et al. notioned that the "context" of a node can be denoted by their co-occurrence in a random walk path [168]. Formally, they put random walkers over networks to record their walking paths, each of which is composed of a chain of nodes that could be considered as a "sentence" of words in a text corpus. More recently, in

order to diversify the neighborhood of a node, Grover & Leskovec presented biased random walkers—a mixture of breadth-first and width-first search procedures—over networks to produce paths of nodes [82]. With node paths generated, both works leveraged the skip-gram architecture in word2vec to model the structural correlations between nodes in a path. In addition, several other methods have been proposed for learning representations in networks [30, 31, 97, 165, 175]. In particular, to learn network embeddings, Tang et al. decomposed a node's context into first- (friends) and second-order (friends' friends) proximity [207], which was further developed into a semi-supervised model PTE for embedding text data [206].

In this work, we further this direction of research by designing the *metapath2vec* and *metapath2vec++* models to capture heterogeneous structural and semantic correlations exhibited from large-scale networks with multiple types of nodes, which can not be handled by previous models, and apply these models on various network mining tasks.

## 7.7 Conclusion

In this chapter, we formally define the representation learning problem in heterogeneous networks in which there exist diverse types of nodes and links. To address the network heterogeneity challenge, we propose the *metapath2vec* and *metapath2vec++* methods. We develop the meta-path guided random walk strategy in a heterogeneous network, which is capable of capturing both the structural and semantic correlations of differently typed nodes and relations. By leveraging this method, we formalize the heterogeneous neighborhood function of a node, enabling the skip-gram based maximization of the network probability in the context of multiple types of nodes. Finally, we achieve the effective and efficient optimization by presenting a heterogeneous negative sampling technique. Extensive experiments demonstrate that the latent feature representations learned by *metapath2vec* and *metapath2vec++* are able

to improve various heterogeneous network mining tasks, such as similarity search, node classification, and clustering. Our results can be naturally applied to real-world applications in heterogeneous academic networks, such as author, venue, and paper search in academic search services.

Future work includes various optimizations and improvements. For example, 1) the *metapath2vec* and *metapath2vec++* models, as is also the case with DeepWalk and node2vec, face the challenge of large intermediate output data when sampling a network into a huge pile of paths, and thus optimizing the sampling space is an important direction; 2) as is also the case with all meta-path based heterogeneous network mining methods, *metapath2vec* and *metapath2vec++* can be further improved by the automatic learning of meaningful meta-paths; and 3) extending the models to incorporate the dynamics of evolving heterogeneous networks.

# CHAPTER 8

# CONCLUSION AND FUTURE DIRECTIONS

In this chapter, we summarize the observations discovered from massive social and information networks, as well as the computational perspectives on addressing challenges that arose from the large scale of complex, networked data. Further, we elaborate on future directions of big network analytics that can build and improve upon them. The overview of this thesis and its future directions is presented in Figure 8.1.

## 8.1 Summary of Contributions

We aim to understand and model the principles that underpin our highly connected world, from individuals, to groups, to societies. We achieve this by studying a collection of more than one hundred large-scale networks in a wide range of domains, including human communication, online social media, scientific collaboration and citation, the Web, and so on. Our work provides novel insights into the interplay of demographics and diversity with network structures, and we leverage that understanding to develop computational models to predict network phenomena.

In particular, the focus of this thesis lies in two components of big network analytics: demographics and diversity. Work in each component is proceeded with a combination of empirical discoveries and measurements from big networks, computational modeling and learning of network problems, and large-scale predictive experiments and applications.

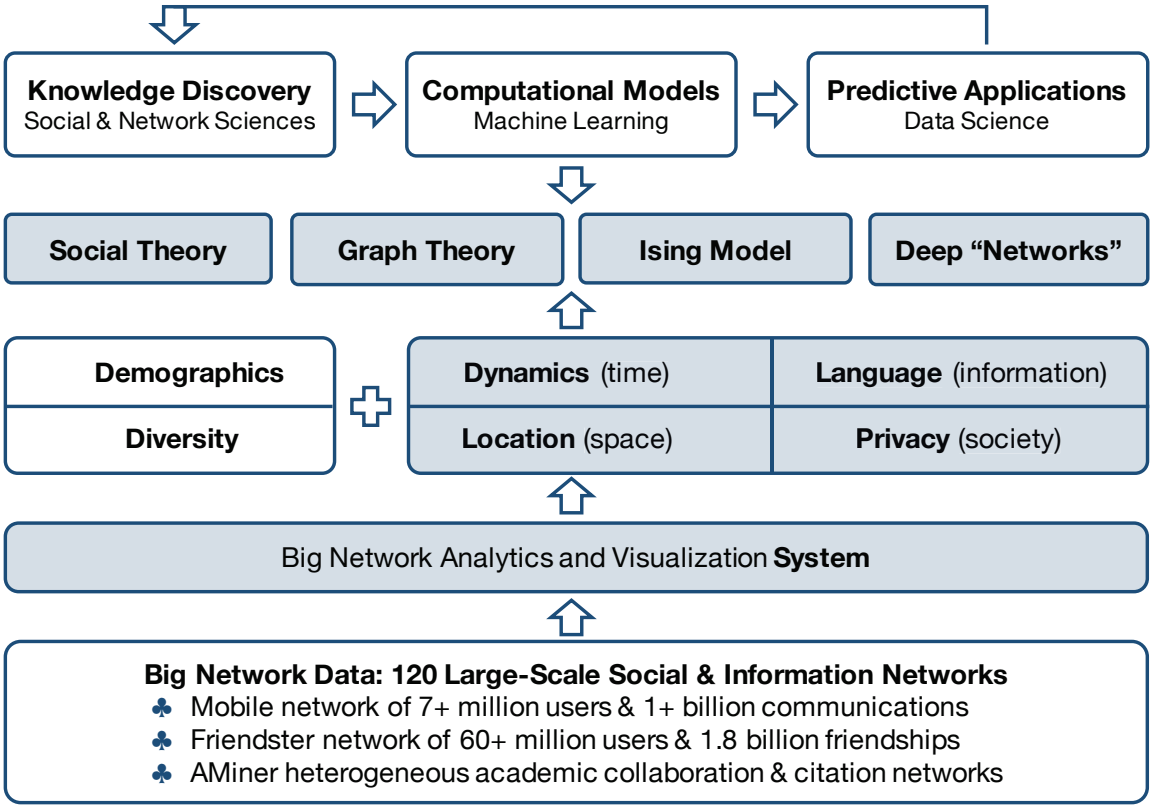



Figure 8.1. The overview of the thesis and future directions. Shading blocks indicate future directions.

In the first part of this thesis, we study both micro- and macro-level network structures that are coupled with user demographics, and provide predictive models of user profiles from networks. First, we discover the evolving patterns of human social strategies, that is, the active expansion of social connections with males and females alike before 35 years of age, and the selective interactions with small, closed, and same-gender social circles after 35. As a consequence, we also find the systematic variations in age-specific small world phenomenon—young people live in the "smallest world" and the most elderly live in the "least small world." Finally, we present factor graph-based computational models that can naturally incorporate structural network features for user demographic prediction. Our extensive experiments demonstrate the predictability of users' gender and age from human communication networks.

In the second part, we investigate network diversity along three dimensions. First, we demonstrate that the structural diversity of the common neighborhood has a significantly different impact on link existence across various networks, leading to the identification of several distinct network superfamilies not discoverable by conventional methods. Further, we find that the topic diversity of a research publication is of surprisingly little relevance to the growth of its authors' scientific impact in academic networks, and, by contrast, that the authors' authority on the publication topic is crucial to impact growth. Finally, we present neural network-based computational models to learn unsupervised latent representations for diverse types of nodes in heterogeneous information networks, advancing conventional heterogeneous network mining and learning tasks, such as node classification, clustering, and similarity search.

## 8.2 Future Directions

With the increasing availability of big network data that are coupled with user behavior along many dimensions, including time, space, information, and society, there are many exciting future directions to explore, such as network dynamics, location based networks, language usage, and privacy. Accordingly, research problems and solutions derived from these network data should be applicable to real-world tasks. More importantly, it will be crucial to advance this field from theoretical and computing perspectives, including the (re)examination and development of network theories and computational models in the context of big networks.

**Network dynamics.** Networks are commonly associated with temporal information. When structures meet with time, it is straightforward to ask how networks evolve over time [121]. For example, Chapter 2 tells us that young females and males have a strong tendency to connect and interact with each other, leading to the following interesting questions: How do their social networking behaviors vary across

different hours of a day, different days of a week, different seasons of a year, and so on? Consequently, how do network structures and properties change over time? Answers to these questions will provide us with a richer understanding of human networking behavior.

**Networks and language.** Although information usually propagates through social networks in the form of natural language, we have limited knowledge about the influence that language has on information diffusion over networks, as well as on the formation and maintenance of social relationships. It would be interesting to characterize the ways in which natural language shapes networks and, reversely, how network structures influence language usage.

**Graph and social theories.** We studied networks, with a focus on social and information networks, which requires domain knowledge from social science, mathematical formulation from graph theory and physics, and computational perspectives from data mining and machine learning. Graph theory provides network science with theoretical foundations, while social theory abstracts empirical observations into principles. It is, as always, vital to supply network analysis and mining with a continuous development of graph and social theories. In particular, we are interested in the interaction area produced at the intersection of graphical and Ising models for future directions.

**Deep "network" models.** Deep learning models can learn latent representations that capture the internal relations from rich, complex data of various modalities, such as image, audio, and language. As a kind of complex data that encodes the chaos, order, and dynamics of human interactions, social and information networks are similarly rich and complex, and may thus also be particularly amenable to deep learning. It would be exciting to study how deep learning techniques can help and potentially shape network science.

**Big network computation and visualization systems.** This thesis focuses on identifying and addressing problems in big networks, wherein computational challenges naturally arise. Currently, there exist several big graph mining platforms, such as SNAP, GraphX, Giraph, GraphChi, etc. Network science would significantly benefit from advances in the computational paradigm, system design and implementation, and graph visualization.

Finally, when looking back to the scientific discovery process in nature, we know that as early as 1808, people started to characterize the existence of atoms [37], which are "the smallest constituent unit of ordinary matter that has the properties of a chemical element [139]." It was nearly one century later, in 1902, when Gilbert Newton Lewis further discovered the covalent bond between atoms, providing an explanation to the phenomenon that matter (e.g., graphite, diamond, fullerene, nanotube, and graphene) composed of the same atoms or elements (carbon atoms) displays markedly different properties. These remarkable differences arise from the different ways in which these atoms are organized by covalent bonds (interactions). Since 1860s, during which Dmitri Mendeleev envisioned the ordered arrangement of all elements into a periodic table [142], we have further realized that the boundless universe is formed by the combinations of around one hundred elements that hold different attributes and traits.

Throughout the course of this thesis, we have been asking an analogous question: "Can society be characterized as a network of 'social atoms', and is it organized under some obvious, yet unknown, mechanisms?" This thesis begins to address these questions by studying the various ways that diverse individuals are embedded in and interact within social and information networks. For the long-term vision, we aim to further explore and discover the underlying organizing principles that drive the formation and evolution of modern society.

# BIBLIOGRAPHY


1. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 2016.

2. D. E. Acuna, S. Allesina, and K. P. Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, Sept. 2012.

3. L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001.

4. A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola. Distributed large-scale natural graph factorization. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*, pages 37–48. ACM, 2013. ISBN 978-1-4503-2035-1.

5. M. Ahmed, S. Spagna, F. Huici, and S. Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of ACM International Conference on Web Search and Data Mining (WSDM '13)*, pages 607–616. ACM, 2013.

6. K. J. Ajrouch, A. Y. Blandon, and T. C. Antonucci. Social networks among men and women: The effects of age and socioeconomic status. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 60(6):S311–S317, 2005.

7. T. Aledavood, E. López, S. G. B. Roberts, F. Reed-Tsochas, E. M. Egido, R. I. M. Dunbar, and J. Saramäki. Channel-specific daily patterns in mobile phone communication. *CoRR*, abs/1507.04596, 2015.

8. L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences (PNAS)*, 97 (21):11149–11152, 2000.

9. T. C. Antonucci and H. Akiyama. Social networks in adult life and a preliminary examination of the convoy model. *Journal of Gerontology*, 42(5):519–527, 1987.

10. N. E. Aristotle and V. Book. Aristotle in 23 volumes, vol. 19, translated by h. rackham, 1934.

11. L. Backstrom and J. M. Kleinberg. Romantic partnerships and the dispersion of social ties: a netwrok analysis of relationship status on facebook. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '14)*, pages 831–841, 2014.

12. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference (WebSci '12)*, pages 33–42. ACM, 2012.

13. A.-L. Barabási. *Network Science*. Cambridge University Press, 2016.

14. A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

15. Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(8):1798–1828, 2013.

16. S. Bethard and D. Jurafsky. Who should I cite: Learning literature search models from citation behavior. In *Proceedings of ACM international conference on Information and knowledge management (CIKM '10)*, pages 609–618. ACM, 2010.

17. B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of International Conference on World Wide Web (WWW '13)*, pages 131–140, 2013.

18. N. Biggs, E. Lloyd, and R. Wilson. Graph theory, 1736-1936, 1986.

19. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res. (JMLR)*, 3:993–1022, 2003.

20. V. D. Blondel, A. Decuyper, and G. Krings. A survey of results on mobile phone datasets analysis. *arXiv preprint arXiv:1502.03406*, 2015.

21. J. Blumenstock and N. Eagle. Mobile divides: Gender, socioeconomic status, and mobile phone use in rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development (ICTD '10)*, pages 6:1–6:10. ACM, 2010.

22. E. Bott. *Family and social network: Roles, norms, and external relationships in ordinary urban families.* Tavistock, 1971.

23. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of International Conference on World Wide Web (WWW7)*, pages 107–117, 1998.

24. R. S. Burt. *Structural Holes : The Social Structure of Competition*. Cambridge, Mass.: Harvard University Press, 1995.

25. R. S. Burt. Measuring age as a structural concept. *Social Networks*, 13(1):1–34, 1991.

26. F. Calabrese, L. Ferrari, and V. Blondel. Urban sensing using mobile phones network data: A survey of research. *ACM Comput. Surv. (CSUR)*, 2014.

27. C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific reports*, 3, 2013.

28. C. Castillo, D. Donato, and A. Gionis. Estimating the number of citations using author reputation. In *Proceedings of International Symposium on String Processing and Information Retrieval (SPIRE '07)*, pages 107–117. Springer, 2007.

29. D. Chakrabarti, S. Funiak, J. Chang, and S. A. Macskassy. Joint inference of multiple label types in large networks. In *Proceedings of International Conference on Machine Learning (ICML '14)*, pages 874–882, 2014.

30. S. Chang, W. Han, J. Tang, G.-J. Qi, C. C. Aggarwal, and T. S. Huang. Heterogeneous network embedding via deep architectures. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '15)*, pages 119–128. ACM, 2015.

31. T. Chen and Y. Sun. Task-guided and path-augmented heterogeneous network embedding for author identification. In *Proceedings of ACM International Conference on Web search and Data Mining (WSDM '17)*, page na. ACM, 2017.

32. J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of International Conference on World Wide Web (WWW '14)*, pages 925–936, 2014.

33. N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

34. S. Comandur. Escape. `https://bitbucket.org/seshadhri/escape`, 2016.

35. B. Cornwell. Age trends in daily social contact patterns. *Research on Aging*, 33 (5):598–631, 2011.

36. B. Cornwell, E. O. Laumann, and L. P. Schumm. The social connectedness of older adults: A national profile. *American Sociological Review*, 73(2):185–203, 2008.

37. J. Dalton. *A new system of chemical philosophy*, volume 1. Rickerstaff, 1808.

38. N. L. Danigelis, M. Hardy, and S. J. Cutler. Population aging, intracohort aging, and sociopolitical attitudes. *American Sociological Review*, 72(5):812–830, 2007.

39. Y. Ding, R. Rousseau, and D. Wolfram. *Measuring scholarly impact: Methods and practice*. Springer, 2014.

40. Y. Ding, G. Zhang, T. Chambers, M. Song, X. Wang, and C. Zhai. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65(9):1820–1833, 2014.

41. P. S. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. *Science*, 301(5634):827–829, 2003.

42. P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '01)*, pages 57–66, 2001.

43. Y. Dong, J. Tang, T. Lou, B. Wu, and N. V. Chawla. How long will she call me? distribution, social theory and duration prediction. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD '13)*, pages 16–31. Springer, 2013.

44. Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*, pages 15–24. ACM, 2014.

45. Y. Dong, R. A. Johnson, and N. V. Chawla. Will this paper increase your $h$-index?: Scientific impact prediction. In *Proceedings of ACM International Conference on Web search and Data Mining (WSDM '15)*, pages 149–158. ACM, 2015.

46. Y. Dong, R. A. Johnson, Y. Yang, and N. V. Chawla. Collaboration signatures reveal scientific impact. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '15)*, pages 480–487, 2015.

47. Y. Dong, F. Pinelli, Y. Gkoufas, Z. Nabi, F. Calabrese, and N. V. Chawla. Inferring unusual crowd events from mobile phone call detail records. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD '15)*, pages 474–492. Springer International Publishing, 2015.

48. Y. Dong, J. Tang, N. V. Chawla, T. Lou, Y. Yang, and B. Wang. Inferring social status and rich club effects in enterprise communication networks. *PLoS ONE*, 10:e0119446, 03 2015.

49. Y. Dong, J. Zhang, J. Tang, N. V. Chawla, and B. Wang. Coupledlp: Link prediction in coupled networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '15)*, pages 199–208. ACM, 2015.

50. Y. Dong, R. A. Johnson, and N. V. Chawla. Can scientific impact be predicted? *IEEE Transactions on Big Data (TBD)*, 2(1):18–30, 2016.

51. Y. Dong, R. A. Johnson, J. Xu, and N. V. Chawla. Structural diversity and homophily: A study across more than one hundred large-scale networks. *CoRR*, abs/1602.07048, 2016. URL `http://arxiv.org/abs/1602.07048`.

52. Y. Dong, O. Lizardo, and N. V. Chawla. Do the young live in a "smaller world" than the old? age-specific degrees of separation in a large-scale mobile communication network. *CoRR*, abs/1606.07556, 2016. URL `http://arxiv.org/abs/1606.07556`.

53. Y. Dong, N. V. Chawla, J. Tang, Y. Yang, and Y. Yang. User modeling on demographic attributes in big mobile social networks. *ACM Transactions on Information Systems (TOIS)*, 2017 (accepted).

54. N. Du, C. Faloutsos, B. Wang, and L. Akoglu. Large human communication networks: Patterns and a utility-driven generator. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, pages 269–278. ACM, 2009.

55. N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106(36):15274–15278, 2009.

56. N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, May 2010.

57. D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

58. M. Ercsey-Ravasz, R. N. Lichtenwalter, N. V. Chawla, and Z. Toroczkai. Range-limited centrality measures in complex networks. *Phys. Rev. E*, 85:066103, Jun 2012.

59. P. Erdős and A. Rényi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

60. P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

61. P. L. Erdös, I. Miklós, and Z. Toroczkai. New classes of degree sequences with fast mixing swap markov chain sampling. *CoRR*, abs/1601.08224, 2016. URL `http://arxiv.org/abs/1601.08224`.

62. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review (SIGCOMM '99)*, pages 251–262, 1999.

63. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9 (Aug):1871–1874, 2008.

64. Z. Fang, X. Zhou, J. Tang, W. Shao, A. Fong, L. Sun, Y. Ding, L. Zhou, and J. Luo. Modeling paying behavior in game social networks. In *Proceedings of ACM international conference on Information and knowledge management (CIKM '14)*, pages 411–420, 2014.

65. O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

66. R. Frankham. Genetics and extinction. *Biological conservation*, 126(2):131–140, 2005.

67. L. C. Freeman. Centered graphs and the structure of ego networks. *Mathematical Social Sciences*, 3(3):291–304, 1982.

68. H. Gao, J. Tang, X. Hu, and H. Liu. Modeling temporal effects of human mobile behavior on location-based social networks. In *Proceedings of ACM international conference on Information and knowledge management (CIKM '13)*, pages 1673–1678, 2013.

69. E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, 1955.

70. J. Gehrke, P. Ginsparg, and J. M. Kleinberg. Overview of the 2003 KDD Cup. *SIGKDD Explorations*, 5(2):149–151, 2003.

71. A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

72. D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia (HYPERTEXT '98)*, pages 225–234. ACM, 1998.

73. E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4): 1141–1144, 1959.

74. M. Gladwell. The coolhunt: who decides whats cool? certain kids in certain places–and only the coolhunters know who they are. *The New Yorker*, 17, 1997.

75. M. Gladwell. *The tipping point: How little things can make a big difference*. Little, Brown, 2006.

76. Y. Goldberg and O. Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722, 2014.

77. M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

78. W. J. Goode. *World revolution and family patterns.* Free Press Glencoe, 1963.

79. M. Granovetter. Problems of explanation in economic sociology. *Networks and organizations: Structure, form, and action*, 25:56, 1992.

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

81. M. Granovetter. Economic action and social structure: The problem of embeddedness. *The American Journal of Sociology*, 1985.

82. A. Grover and J. Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '16)*, pages 855–864. ACM, 2016.

83. S. Hakami. On the realizability of a set of integers as degrees of the vertices of a graph. *SIAM Journal Applied Mathematics*, 1962.

84. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

85. V. Havel. A remark on the existence of finite graphs. *Casopis Pest. Mat*, 80 (477-480):1253, 1955.

86. F. Heider. *The psychology of interpersonal relations.* Wiley, 1958.

87. K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li. Rolx: structural role extraction & mining in large graphs. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*, pages 1231–1239. ACM, 2012.

88. S. C. Herring. Gender and power in on-line communication. In *Handbook of language and gender*, page 202. Wiley-Blackwell, 2003.

89. C. A. Hidalgo and C. Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017 – 3024, 2008.

90. J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences (PNAS)*, 102(46):16569–16572, 2005.

91. P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97 (460):1090–1098, 2002.

92. P. Holme, C. R. Edling, and F. Liljeros. Structure and time evolution of an internet dating community. *Social Networks*, 26(2):155–174, 2004.

93. L. Hong, A. S. Doumith, and B. D. Davison. Co-factorization machines: Modeling user interests and predicting individual decisions in twitter. In *Proceedings of ACM International Conference on Web search and Data Mining (WSDM '13)*, pages 557–566. ACM, 2013.

94. J. E. Hopcroft, T. Lou, and J. Tang. Who will follow you back? reciprocal relationship prediction. In *Proceedings of ACM international conference on Information and knowledge management (CIKM '11)*, 2011.

95. J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of International Conference on World Wide Web (WWW '07)*, pages 151–160, 2007.

96. X. Hu and H. Liu. Social status and role analysis of Palin's email network. In *Proceedings of International Conference on World Wide Web (WWW '12 Companion)*, pages 531–532. ACM, 2012.

97. X. Huang, J. Li, and X. Hu. Label informed attributed network embedding. In *Proceedings of ACM International Conference on Web search and Data Mining (WSDM '17)*, page na. ACM, 2017.

98. M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11)*, pages 1298–1306. ACM, 2011.

99. E. M. Jin, M. Girvan, and M. E. J. Newman. Structure of growing social networks. *Phys. Rev. E*, 64:046132, 2001.

100. E. Katz and P. F. Lazarsfeld. *Personal Influence*. The Free Press, New York, USA, 1955.

101. D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03)*, pages 137–146, 2003.

102. I. King, M. R. Lyu, and H. Ma. Introduction to social recommendation. In *Proceedings of International Conference on World Wide Web (WWW '10)*, pages 1355–1356. ACM, 2010.

103. J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing (STOC '00)*, pages 163–170. ACM, 2000.

104. J. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

105. J. Kleinberg. Small-world phenomena and the dynamics of information. *Advances in neural information processing systems (NIPS '02)*, 1:431–438, 2002.

106. J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

107. J. Kleinberg and S. Oren. Mechanisms for (mis)allocating scientific credit. In *Proceedings of ACM symposium on Theory of computing (STOC '11)*, pages 529–538. ACM, 2011.

108. X. Kong and P. S. Yu. Semi-supervised feature selection for graph classification. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*, pages 793–802, 2010.

109. Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*, pages 426–434. ACM, 2008.

110. L. Kovanen, K. Kaski, J. Kertész, and J. Saramäki. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences (PNAS)*, 110(45):18070–18075, 2013.

111. D. Krackhardt. *The Strength of Strong ties*. Cambridge, Harvard Business School Press, Hershey, USA, 1992.

112. F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory (TOIT)*, 47: 498–519, 2001.

113. S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Machematical Statistics*, 22(1):79–86, 1951.

114. J. Kunegis. Konect: the koblenz network collection. In *Proceedings of International Conference on World Wide Web (WWW '13 Companion)*, pages 1343–1350, 2013.

115. P. F. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society, New York: Van Nostrand*, pages 8–66, 1954.

116. P. F. Lazarsfeld, B. Berelson, and H. Gaudet. *The people's choice: How the voter makes up his mind in a presidential campaign*. Columbia University Press, New York, USA, 1944.

117. D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science*, 323(5915):721, 2009.

118. F. Le Play. *Frederic Le Play on Family, Work, and Social Change*. University of Chicago Press, Chicago, 1982.

119. Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

120. J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of International Conference on World Wide Web (WWW '08)*, pages 915–924. ACM, 2008.

121. J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, pages 177–187. ACM, 2005.

122. J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of International Conference on World Wide Web (WWW '08)*, pages 695–704. ACM, 2008.

123. J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res. (JMLR)*, 11(Feb):985–1042, 2010.

124. Y. Li. Toward A qualitative search engine. *IEEE Internet Computing*, 2(4): 24–29, 1998.

125. D. Liben-Nowell and J. M. Kleinberg. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

126. L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *Proceedings of ACM international conference on Information and knowledge management (CIKM '10)*, pages 199–208. ACM, 2010.

127. O. Lizardo. How cultural tastes shape personal networks. *American Sociological Review*, 71(5):778–807, 2006.

128. Q. Llimona, J. Luque, X. Anguera, Z. Hidalgo, S. Park, and N. Oliver. Effect of gender and call duration on customer satisfaction in call center big data. In *INTERSPEECH '15*, 2015.

129. T. Lou and J. Tang. Mining structural hole spanners through information diffusion in social networks. In *Proceedings of International Conference on World Wide Web (WWW '13)*, pages 825–836, 2013.

130. T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding. Learning to predict reciprocity and triadic closure in social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(2):5:1–5:25, 2013.

131. H. Ma. On measuring social friend interest similarities in recommender systems. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*, pages 465–474. ACM, 2014.

132. H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *Proceedings of ACM International Conference on Web search and Data Mining (WSDM '11)*, pages 287–296, 2011.

133. L. Magee. $R^2$ measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, 44(3):250–253, 1990.

134. A. M. Manago, T. Taylor, and P. M. Greenfield. Me and my 400 friends: the anatomy of college students' facebook networks, their communication patterns, and well-being. *Developmental Psychology*, 48(2):369, 2012.

135. C. S. Marcum. Age differences in daily social activities. *Research on Aging*, 35 (5):612–640, 2013.

136. P. V. Marsden. Core discussion networks of americans. *American Sociological Review*, pages 122–131, 1987.

137. P. V. Marsden and K. E. Campbell. Measuring tie strength. *Social Forces*, 63 (2):482–501, 1984.

138. M. Max-Neef, A. Elizalde, and M. Hopenhayn. Development and human needs. *Real-life economics: Understanding wealth creation*, pages 197–213, 1992.

139. A. D. McNaught and A. D. McNaught. *Compendium of chemical terminology*, volume 1669. Blackwell Science Oxford, 1997.

140. M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444, 2001.

141. M. Mead. *Culture and commitment: a study of the generation gap.* Natural History Press, 1970.

142. D. Mendelejeff. Ueber die beziehungen der eigenschaften zu den atomgewichten der elemente. *Z. Chem*, 12:405–406, 1869.

143. L. Meng, Y. Hulovatyy, A. Striegel, and T. Milenković. On the interplay between individuals' evolving interaction patterns and traits in dynamic multiplex social networks. *IEEE Transactions on Network Science and Engineering*, 3(1):32–43, 2016.

144. M. Michelson and S. A. Macskassy. What blogs tell us about websites: A demographics study. In *Proceedings of ACM International Conference on Web search and Data Mining (WSDM '11)*, pages 365–374. ACM, 2011.

145. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL `http://arxiv.org/abs/1301.3781`.

146. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS '13)*, pages 3111–3119, 2013.

147. S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

148. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303 (5663):1538–1542, March 2004.

149. G. Miritello, R. Lara, M. Cebrian, and E. Moro. Limited communication capacity unveils strategies for human interaction. *Scientific Reports*, 3, 2013.

150. A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (IMC'07)*, pages 29–42, 2007.

151. A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the demographics of twitter users. In *International AAAI Conference on Web and Social Media (ICWSM '11)*, 2011.

152. K. Mo, B. Tan, E. Zhong, and Q. Yang. Your phone understands you. In *Nokia MDC '12*, 2012.

153. J. Moody. Peer influence groups: identifying dense clusters in large networks. *Social Networks*, 23(4):261–283, 2001.

154. C. Moore and M. E. Newman. Epidemics and percolation in small-world networks. *Phys. Rev. E*, 61(5):5678, 2000.

155. J. L. Moreno. *Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama.* Beacon House, 1953.

156. A. Munch, J. M. McPherson, and L. Smith-Lovin. Gender, children, and social contact: The effects of childrearing for men and women. *American Sociological Review*, pages 509–520, 1997.

157. G. P. Murdock. Kin term patterns and their distribution. *Ethnology*, 9(2): 165–208, 1970.

158. K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (UAI '99)*, pages 467–475, 1999.

159. National Research Council. *Network Science*. The National Academies Press, 2005.

160. J. Neville and D. Jensen. Leveraging relational autocorrelation with latent group models. In *Proceedings of the 4th international workshop on Multi-relational mining*, pages 49–55. ACM, 2005.

161. M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64(2):025102, 2001.

162. M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006.

163. M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, 2001.

164. J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences (PNAS)*, 2007.

165. M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 1105–1114. ACM, 2016.

166. V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. M. Dunbar. Sex differences in intimate relationships. *Scientific Reports*, 2:370, 2012.

167. T. Parsons. Age and sex in the social structure of the united states. *American Sociological Review*, pages 604–616, 1942.

168. B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*, pages 701–710. ACM, 2014.

169. K. W. Phillips. How diversity makes us smarter. *Scientific American*, 311(4), 2014.

170. H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of ACM International Conference on Web search and Data Mining (WSDM '13)*, pages 365–374. ACM, 2013.

171. F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences (PNAS)*, 105(45):17268–17272, 2008.

172. S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage, 2002.

173. S. Redner. On the meaning of the h -index. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(03):L03005, 2010.

174. X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han. Cluscite: Effective citation recommendation by information network-based clustering. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*, 2014.

175. X. Ren, W. He, M. Qu, C. R. Voss, H. Ji, and J. Han. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *KDD '16*, pages 1825–1834. ACM, 2016.

176. X. Rong. word2vec parameter learning explained. *CoRR*, abs/1411.2738, 2014. URL `http://arxiv.org/abs/1411.2738`.

177. R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI' 15*, pages 4292–4293, 2015. URL `http://networkrepository.com`.

178. S. Ruggles. The decline of intergenerational coresidence in the united states, 1850 to 2000. *American Sociological Review*, 72(6):964–989, 2007.

179. J. Saramäki and E. M. Egido. From seconds to months: multi-scale dynamics of mobile telephone calls. *CoRR*, abs/1504.01479, 2015. URL `http://arxiv.org/abs/1504.01479`.

180. M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskovec. Mobile call graphs: beyond power-law and lognormal distributions. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*, pages 596–604. ACM, 2008.

181. H. Shen, D. Wang, C. Song, and A.-L. Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. In *AAAI'14*, pages 291–297, 2014.

182. H.-W. Shen and A.-L. Barabási. Collective credit allocation in science. *Proceedings of the National Academy of Sciences (PNAS)*, 111(34):12325–12330, 2014.

183. X. Shi, L. A. Adamic, and M. J. Strauss. Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 378(1):33–47, 2007.

184. X. Shi, L. A. Adamic, B. L. Tseng, and G. S. Clarkson. The impact of boundary spanning scholarly publications and patents. *PLoS ONE*, 4(8):e6547, 08 2009.

185. X. Shi, B. Tseng, and L. Adamic. Information diffusion in computer science citation networks. In *International AAAI Conference on Web and Social Media (ICWSM '09)*, 2009.

186. X. Shi, J. Leskovec, and D. A. McFarland. Citing for high impact. In *Proceedings of the 10th annual joint conference on Digital libraries (JCDL '10)*, pages 49–58. ACM, 2010.

187. B.-E. Shie, S. Y. Philip, and V. S. Tseng. Mining interesting user behavior patterns in mobile commerce environments. *Applied intelligence*, 38(3):418–435, 2013.

188. R. Shields. Cultural topology: The seven bridges of königsburg, 1736. *Theory, Culture & Society*, 29(4-5):43–57, 2012.

189. E. J. Smith, C. S. Marcum, A. Boessen, Z. W. Almquist, J. R. Hipp, N. N. Nagle, and C. T. Butts. The relationship of age to personal network size, relational multiplexity, and proximity to alters in the western united states. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 70(1):91–99, 2015.

190. Z. Smoreda and C. Licoppe. Gender-Specific Use of the Domestic Telephone. *Social Psychology Quarterly*, 63(3):238–252, 2000.

191. R. Sosic and J. Leskovec. Large scale network analytics with snap. In *Proceedings of International Conference on World Wide Web (WWW '15 Companion)*, pages 1537–1538. ACM, 2015.

192. B. Spencer. Mobile users can't leave their phone alone for six minutes and check it up to 150 times a day. `http://www.dailymail.co.uk/news/article-2276752/Mobile-users-leave-phone-minutes-check-150-times-day.html`, 2013.

193. R. C. Sprinthall. *Basic statistical analysis*. 2011.

194. A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, J. E. Larsen, and S. Lehmann. Measuring large-scale social networks with high resolution. *PLoS One*, 9(4):e95978, 2014.

195. M. Strathern. Improving ratings: audit in the British university system. *European Review*, 5(03):305–321, 1997.

196. Y. Sun and J. Han. *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers, 2012.

197. Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, pages 797–806. ACM, 2009.

198. Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proceedings of the VLDB Endowment (VLDB '11)*, pages 992–1003, 2011.

199. Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla. When will it happen?: Relationship prediction in heterogeneous information networks. In *Proceedings of ACM International Conference on Web search and Data Mining (WSDM '12)*, pages 663–672. ACM, 2012.

200. Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*, pages 1348–1356. ACM, 2012.

201. M. Szell and S. Thurner. How women organize social networks different from men. *Scientific Reports*, 3, July 2013.

202. J. Tang and J. Zhang. A discriminative approach to topic-based citation recommendation. *Advances in Knowledge Discovery and Data Mining*, pages 572–579, 2009.

203. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*, pages 990–998, 2008.

204. J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, pages 807–816, 2009.

205. J. Tang, S. Wu, and J. Sun. Confluence: Conformity influence in large social networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13)*, pages 347–355. ACM, 2013.

206. J. Tang, M. Qu, and Q. Mei. PTE: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, pages 1165–1174. ACM, 2015. ISBN 978-1-4503-3664-2.

207. J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *Proceedings of International Conference on World Wide Web (WWW '15)*. ACM, 2015.

208. J. Tang, T. Lou, J. Kleinberg, and S. Wu. Transfer learning to infer social ties across heterogeneous networks. *ACM Transactions on Information Systems (TOIS)*, 34(2):7:1–7:43, Apr. 2016.

209. L. Tang and H. Liu. Leveraging social media networks for classification. *Data Mining and Knowledge Discovery (DMKD)*, 23(3):447–478, 2011.

210. L. Tang and H. Liu. Relational learning via latent social dimensions. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, pages 817–826, 2009.

211. Z. Toroczkai and K. E. Bassler. Network dynamics: Jamming is limited in scale-free systems. *Nature*, 428(6984):716–716, 2004.

212. J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.

213. W. T. Tutte. The factorization of linear graphs. *Journal of the London Mathematical Society*, 1(2):107–111, 1947.

214. W. T. Tutte. The factors of graphs. *Canad. J. Math*, 4(3):314–328, 1952.

215. J. Ugander. *Computational Perspectives on Large-scale Social Networks*. PhD thesis, Cornell University, 2014.

216. J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences (PNAS)*, 109(16):5962–5966, 2012.

217. J. Ugander, L. Backstrom, and J. Kleinberg. Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In *Proceedings of International Conference on World Wide Web (WWW '13)*, pages 1307–1318, 2013.

218. J. R. Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM*, 23(1):31–42, 1976.

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

220. B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

221. D. Vu, A. Asuncion, D. Hunter, and P. Smyth. Dynamic egocentric models for citation networks. In *Proceedings of International Conference on Machine Learning (ICML '11)*, pages 857–864, 2011.

222. C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10)*, pages 203–212, 2010.

223. D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabási. Human mobility, social ties, and link prediction. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11)*, pages 1100–1108. ACM, 2011.

224. D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

225. J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

226. D. J. Watts. Computational social science: Exciting progress and future directions. *The Bridge on Frontiers of Engineering*, 43(4):5–10, 2013.

227. D. J. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton University Press, 1999.

228. D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, pages 440–442, Jun 1998.

229. B. Wellman, R. Y.-l. Wong, D. Tindall, and N. Nazer. A decade of network change: Turnover, persistence and stability in personal communities. *Social Networks*, 19(1):27–50, 1997.

230. Wikipedia. Postpaid mobile phone, accessed on jan. 13th, 2017. `https://en.wikipedia.org/wiki/Postpaid_mobile_phone`, .

231. Wikipedia. Prepay mobile phone, accessed on jan. 13th, 2017. `https://en.wikipedia.org/wiki/Prepay_mobile_phone`, .

232. R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *Proceedings of International Conference on World Wide Web (WWW '10)*, pages 981–990, 2010.

233. R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: Learning to estimate future citations for literature. In *Proceedings of ACM international conference on Information and knowledge management (CIKM '11)*, pages 1247–1252. ACM, 2011.

234. R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. In *Proceedings of annual joint conference on Digital libraries (JCDL '12)*, pages 51–60. ACM, 2012.

235. S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(1), 2007.

236. J. Ying, Y.-J. Chang, C.-M. Huang, and V. S. Tseng. Demographic prediction based on user's mobile behaviors. In *Nokia MDC '12*, 2012.

237. X. Yu, Q. Gu, M. Zhou, and J. Han. Citation prediction in heterogeneous bibliographic networks. In *Proceedings of the 2012 SIAM International Conference on Data Mining (SDM '12)*, pages 1119–1130, 2012.

238. R. Zafarani and H. Liu. Social computing data repository at ASU, 2009. URL `http://socialcomputing.asu.edu`.

239. J. Zhang, J. Tang, and J. Li. Expert finding in a social network. In *Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA '07)*, pages 1066–1069, 2007.

240. J. Zhang, J. Tang, C. Ma, H. Tong, Y. Jing, and J. Li. Panther: Fast top-k similarity search on large networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '15)*, pages 1445–1454. ACM, 2015.

241. Y. Zhao, G. Wang, P. S. Yu, S. Liu, and S. Zhang. Inferring social roles and statuses in social networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13)*, pages 695–703, 2013.

242. Y. Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29, 2015.

*This document was prepared & typeset with* pdfLaTeX*, and formatted with* NDdiss2$_\varepsilon$ *classfile (v3.2013[2013/04/16]) provided by Sameer Vijay and updated by Megan Patnott.*