The Structure of Social Networks: Modeling, Sampling, and Inference

Naghmeh Momeni Taramsari

Doctor of Philosophy

Department of Electrical and Computer Engineering

McGill University

Montreal,Quebec

April 2017

A thesis submitted to McGill University
in partial fulfillment of the requirements of
the degree of Doctor of Philosophy

TABLE OF CONTENTS

ABSTRACT vi

ABRÉGÉ vii

ACKNOWLEDGEMENTS viii

PREFACE AND CONTRIBUTION OF AUTHORS ix

LIST OF FIGURES x

LIST OF TABLES xiii

1 Introduction 1

1.1 Network Perspective 1
1.2 Notation and Terminology 3
1.3 Historical Background on Network Science 5
1.4 Social Networks and Social Outcomes 7
1.5 Thesis Outline 9
1.6 Publications 12

2 Network Sampling 13

2.1 Chapter Outline 13

2.2 Introduction . . . 13
2.3 Challenges of Sampling Offline Social Networks . . . 15
2.4 Fixed-choice Designs . . . 16
2.5 Chapter Summary . . . 17

3 Paper: Statistical Inference for Fixed-choice Design Incorporating Strong and Weak Ties . . . 18

4 Additional Simulation Results for Network Sampling . . . 52

4.1 Chapter Outline . . . 52
4.2 Introduction . . . 52
4.3 Transitivity . . . 53
4.4 Community Structure . . . 53
4.5 Popularity Bias . . . 54
4.6 Chapter Summary . . . 58

5 Introduction to Alter Sampling . . . 59

5.1 Chapter Outline . . . 59
5.2 Introduction . . . 59
5.3 Applications of Alter Sampling . . . 60
5.3.1 Acquaintance Vaccination . . . 60
5.3.2 Health Monitoring . . . 62
5.3.3 Information Diffusion on Social Media . . . 63
5.3.4 Early Detection of Natural Disasters . . . 64
5.3.5 Health Intervention . . . 64
5.4 Chapter Summary . . . 65

6 Paper: Effectiveness of Alter Sampling in Various Social Networks . . . 66

7 Alter Sampling and the Friendship Paradox . . . 86

7.1 Chapter Outline . . . 86
7.2 Introduction . . . 86
7.3 The Friendship Paradox . . . 87
7.4 Introducing the Generalized Friendship Paradox . . . 90
7.5 Chapter Summary . . . 92

8 Paper: Friendship Paradox and the Inequalities in Social Networks . . . 94

9 Paper: Generalized Friendship Paradox in Growing Networks . . . 120

10 Paper: Quantifying and Measuring the Friendship Paradox . . . 139

11 Future Work 154

11.1 Multiplex Data 154
11.2 Additional Socio-centric Data 154
11.3 Heterogeneous Tie Strength 155
11.4 Effect of Inequalities in Social Networks on Subjective Well-being . 156
11.5 Interplay between Structural and Non-structural Inequality 156

REFERENCES 158

ABSTRACT

Social networks are essential tools for modeling social dynamics. Their structure affects and is affected by the behavior of individuals that constitute them. Many studies have related the structure of social networks to various social and individual outcomes. In many studies, the first step towards network analysis is to observe the network. If the full network is infeasible to acquire, network sampling methods are employed. Sampling offline social networks involves interviewing people. Since respondent fatigue is a pressing problem, standard practice is to ask each respondent only a limited number of names. This throws away much information about the network structure. In this thesis, we focus on the problem of estimating the structural properties of the original social network from such survey data. We provide reliable estimators that incorporate link heterogeneity.

We then focus on applications where knowledge over the global structure of the social network is unfeasible, and efficient methods are needed to identify nodes with certain properties without having to sample the network. We focus on a method called Alter Sampling, which was originally introduced in network epidemiology. We demonstrate its effectiveness in various social networks with different structural properties. Then we highlight insights that this ubiquitous effectiveness provides about how social networks are organized. We discuss the relations to the so-called Friendship Paradox and its generalized version, and provide metrics to quantify how local structural and non-structural properties of nodes compare with their neighbors.

ABRÉGÉ

La modélisation des dynamiques sociales dépend étroitement sur la structure des réseaux sociaux qui influe sur et est également influencée par le comportement des individus qui composent le réseau. De nombreuses études constatent qu'il existe des liens serrés entre les réseaux sociaux et divers résultats sur non seulement le plan individuel, mais aussi le plan social. En générale, l'observation du réseau est la première étape de son analyse. Par la suite, si la consitution du réseau complet est impossible à déterminer, les méthodes d'échantillonage en réseaux sont souvent employées. Pour réaliser un échantillonge des réseaux sociaux hors lignes et pour réduire l'effet de la fatigue sur les personnes interrogées, des entretiens sont effectuées desquelles la pratique habituelle est de demander un nombre limité de noms, ce qui gâche une bonne partie de l'information sur la structure sous-jacente du réseau. Dans cette thèse, nous nous concentrons sur comment estimer les caractéristiques structurelles du réseau social original à partir de telles données. Ainsi, nous fournissons des estimateurs fiables qui intègrent l'hétérogénéité des liens.

Motivés par le besoin de développer des méthodes efficaces pour identifier des noeuds ayant certaines caractéristiques sans devoir échantilloner le réseau au complet, nous nous avons ensuite tournés vers les réseaux sociaux pour lesquels il est impossible de cerner leur structure globale. Nous nous sommes penchés sur la méthode «d'échantillonage altérée»(ou Alter Sampling en anglais), qui a d'abord été introduit dans le domaine de l'épidémiologie des réseaux, et nous montrons son efficacité dans divers réseaux dont les caractéristiques structurales sont toutes différentes. Finalement, nous soulignons comment l'efficacité de l'échantillonage altérée nous informe sur l'organisation sociale de ces mêmes réseaux. Nous élaborons sur les relations entre le Paradoxe de l'amitié et sa généralisation, et nous proposons des indicateurs pour quantifier les caractéristiques (non-)structurales locales par rapport à leurs voisins.

ACKNOWLEDGEMENTS

I express my gratitude towards my adviser Prof. Michael Rabbat for his help and support during my PhD. I thank Prof. Sam Harper for agreeing to be on my PhD committee. I would like to thank Prof. Yannis Psaromiligkos for both being on my PhD committee and kindly agreeing to be the internal examiner of the dissertation. I also thank Prof. June Zhang for kindly agreeing to be the external examiner. I thank the feedback and input of the committee during the different stages of my PhD.

Moreover, I would like to wholeheartedly express my sincere gratitude to my dear parents and my dear brother for their love and care throughout life, and to my beloved husband for his love and support during the life we have shared.

PREFACE AND CONTRIBUTION OF AUTHORS

This dissertation is an original intellectual product of the author, Naghmeh Momeni Taramsari. The work presented in this dissertation is the result of research conducted between January 2014 and March 2017 at the department of electrical and computer engineering, McGill university, under supervision of Prof. Michael Rabbat. The following manuscripts have been published in peer-reviewed journals and conferences based on the presented work.

1. N. Momeni, M. Rabbat, "Qualities and Inequalities in Online Social Networks through the Lens of the Generalized Friendship Paradox", PloS one 11.2 (2016): e0143633.
2. N. Momeni, M. Rabbat, "Measuring the Generalized Friendship Paradox in Networks with Quality-Dependent Connectivity", Complex Networks VI. Springer International Publishing, 2015. 45-55.
3. N. Momeni, M. Rabbat, "Inferring network properties from fixed-choice design with strong and weak ties", IEEE Statistical Signal Processing Workshop (SSP), 2016.
4. N. Momeni, M. Rabbat, "Generalized Friendship Paradox: An Analytical Approach", International Conference on Social Informatics. Springer International Publishing, 2014.
5. N. Momeni, M. Rabbat, "Inferring Structural Characteristics of Networks with Strong and Weak Ties from Fixed-Choice Surveys", accepted to appear in IEEE Transactions on Signal and Information Processing over Networks.
   doi: 10.1109/TSIPN.2017.2731053

LIST OF FIGURES

Figure		page
1–1	Example graph to illustrate how clustering coefficient is calculated. . .	5
1–2	The growth in the relative occurrence of the term 'Social Network Analysis' in the corpus of the literature that Google NGram Viewer has indexed. The third word is intentionally added, because only using 'social network' would also return result that might have been related to online social media, instead of the theoretical field of research which is our purpose. Note that in the past 40 years, the relative occurrence has grown over 50 times.	8
3–1	Schematic illustration of Sampling Setup	24
3–2	Different compositions of open triads	29
3–3	Different compositions of triangles	31
3–4	Illustrative example of triangle and open triad after sampling	32
3–5	Distribution of relative error for the approximation made in calculating $E\{m_0^W\}$	36
3–6	Performance of the estimators as a function of N	38

3–7 Performance of estimators for different q and B 40

3–8 Comparing the performance to that of the single-layer collapsed method 41

3–9 Performance of the estimators for the clustering coefficient 42

3–10 Performance of the estimator for real offline network data set 43

3–11 Standard deviation of estimators via jackknife 45

3–12 All possible ways a triangle could be observed 47

3–13 All possible ways an open triad could be observed 48

4–1 Results for transitivity-driven response procedure 55

4–2 Results for community-driven response procedure 56

4–3 Results for popularity-driven response procedure 57

6–1 Empirical distribution of nodes with gain higher than one in directed networks 72

6–2 Empirical distribution of nodes with gain higher than one in undirected networks 73

6–3 Distribution of gain for all networks 74

6–4 The gain of the estimator as a function of in-degree percentile 75

6–5 Performance of estimator for the four network families 77

8–1 Distribution of different nodal attributes 104

8–2 Proportion of neighbor superiority of different types for different nodal attributes 109

8–3 Proportion of neighbor superiority of different types for percentiles of nodal attributes 111

8–4 In-degree distribution of followees of nodes as a function of their in-degree 113

9–1 Four instances of quality distributions used in this paper for Bernoulli and for exponential distributions 127

9–2 Network level quality and friendship paradox for Bernoulli degree distribution 128

9–3 Network level quality and friendship paradox for exponential degree distribution 129

10–1 Example quality distributions used in the model 144

10–2 Critical values of quality and degree for Bernoulli and exponential quality distributions 145

10–3 Fraction of nodes in the quality and friendship paradox for exponential quality distribution 148

10–4 Fraction of nodes in the quality and friendship paradox for Bernoulli quality distribution 149

LIST OF TABLES

Table		page
3–1	Notation used for statistics of the original (unknown) and sampled (observed) networks	25
3–2	Approximate probabilities for response outcome of seeds belonging to triangles or open triads	46
3–3	Probabilities of observing different triangles	47
3–4	Probabilities of observing open triads from triangles	48
3–5	Probabilities of observing open triads from open triads	49
6–1	Network statistics for the directed networks in the data set	72
6–2	Network statistics for the undirected networks in the data set	73
8–1	Summary statistics of the 8 nodal attributes used	103
8–2	Correlation coefficient between nodal attributes	105
8–3	Fraction of nodes experiencing different types of neighbor superiority	106
8–4	Critical values for different types of neighbor superiority	106

CHAPTER 1

Introduction

1.1 Network Perspective

Network science is the field of studying interacting systems via the mathematical analysis of their network representations. The system can be any in which individual parts can be characterized as units that are linked to one another. Units can be, for example, computers connected, cell phones communicating via a cellular network, neurons connected via synapses, humans having social interaction (face-to-face or on social media), banks with transaction flows between them, scientific papers citing one another, airports exchanging travel flows, or web pages connected via hyperlinks. Traditionally, each discipline has studied (1) how these individual units work, and (2) how these units interact. The network framework offers new insights obtained by looking at the *patterns* of connections. For example, since the early 1900s, neurology and neurobiology have made a remarkable progress in studying the nervous system of various species and finding out (1) how neurons work (e.g., their anatomy, polarity, and function), and (2) how neurons interact (for example, what chemicals are used in chemical synopsis or how voltage patterns change in neurons' gap junction during electrical synopsis). The network approach (with a rich literature in neuroscience) looks at how the patterns of structural and functional connection (i.e., the structure of the *neural network*) affect the collective behavior of the system. For example, it has been found that on average,

there are more supra-tentorial inter-hemisphere connections in the female brain than the male brain, and more intra-hemisphere connections in the male brain than the female brain [ISP+14]. This gives on average a stronger ability to the female brains for communication between analytical and intuitive processing, and on average a stronger ability to the male brain to coordinate perception with action [ISP+14]. So, with identical units (neurons), different patterns of connection can lead to different collective outcomes.

Another example can be found in daily life. There is a vast literature of psychology on (1) the internal mechanisms behind human behavior and emotions, and on (2) how people perceive each other's actions and how they judge and react. The network approach offers insight on how the patterns of connections between humans affect collective outcomes. Strictly-hierarchical chain-of-command organizations have completely different outcomes than more flat and fluid organizational structures that are becoming more prevalent in modern management [Ben11]. In society, life in dense tightly-knit communities is a completely different experience than in communities that are more open and homogeneously linked. The former is a patchy-looking society with strong social control and group conformity within clusters [Col88] but low collective cooperation and general trust [Put95, Fuk01].

It is clear that the effect of patterns of connections on system outcomes is not limited to these two examples, and is in fact prevalent in many networked systems. A nice illustrative analogy is given in [Chr10]: organizing carbon atoms in different ways can give us graphite (pencil) or diamond. In both cases, the units are the same atoms, but it is the way they are structured that gives rise to significantly different properties.

Since social networks are important tools for studying various human-related phenomena, in many studies we need methods to observe and measure them, so that we can incorporate them in analyses. The more thoroughly we know the network structure, the more accurate the consequent analysis and predictions would be. In principle, we first transform the links from abstract mathematical entities to measurable quantities. The 'relations' that the links model should be defined quantitatively. For example, in the case of friendship, we need to first quantitatively define what we mean by friendship, so that we can design a procedure to survey friendships to build a network. Defining relations is relatively straightforward in most studies, because each field has developed its conventions.

The crucial step however, is to actually *observe* these relations to build the networks. In an ideal case, after defining the relations, we would take all the nodes and observe all the links, giving us complete knowledge over the network. This rarely happens in practice. For most real networks, that would be highly impractical. Thus we have to devise economical methods to 'infer' the needed structural parameters from a limited set of observations. This is the task of the field of *network sampling*. Network sampling is an integral part of many social network studies. Sometimes we perform sampling to estimate the network structure, and sometimes to find nodes with certain properties. This thesis focuses on these two distinct cases separately, in that order.

In the remainder of this chapter, we provide a brief overview of example applications of the network framework in different disciplines. We discuss why networks are valuable tools to analyze such systems. We emphasize more on social networks. Before all of this, we introduce notation and terminology.

1.2 Notation and Terminology

The network framework models a system of units as a graph. Each unit is modeled as a dot, which is called a *vertex*, or a *node*. Two units are connected via a line if they are related, where the definition of what it means to be related depends on the context of the study. These lines are called *edges*, *links*, or *ties*. These links can have directions or can be undirected. In a directed graph, node y is called the *out-neighbor* of node x if there exists a link that goes from node x to node y. In that case, node x is called an *in-neighbor* of node y. In an undirected graph, if there is a link between nodes x and y, then they are said to be *neighbors* of each other, also sometimes said to *adjacent* to each other. The links can have numbers assigned to them, which characterize some quality of the relation between nodes, depending on the context. These are called *weights*, and such a network is called a *weighted* network. Otherwise the network is called *unweighted*.

In the specific case of *social* networks (to be discussed in Section 1.4), each node is called an *ego*, the neighbors of the ego are called its *alters* (the word *ego* in Latin means 'I', and the word alter means 'other'). In social networks, a link is called a *tie*.

In directed networks, the number of in-neighbors a node has is called its *in-degree*, and similarly, the number of its out-neighbors is called its *out-degree*. In undirected

networks, the number of neighbors of a node is called the *degree* of that node. For weighted networks, the sum of of weights attached links belonging to a node is called the *strength* of the node. We can use the degrees of all nodes to construct the degree distribution of the network, $p(k)$. So $p(k)$ is the fraction of nodes in the network who have degree k, and $\sum_k p(k) = 1$. The first moment of the degree distribution is called the *average degree*, which we denote by $\overline{k}$. We denote the number of nodes in the network by N and the number of links by L.

The connections among nodes are characterized by the *adjacency matrix* A. If we label the nodes $1, 2, \ldots, N$, then the entry A_{ij} is equal to the weight of the link that connects node i to node j. So for undirected networks, A is a symmetric matrix. For unweighted networks, A is a binary matrix: A_{ij} is equal to 1 if nodes i and j are connected, and is 0 otherwise.

The *density* of a network is the ratio of the number of links to the number of all possible links. The latter is equal to $\binom{N}{2}$. So network density ρ is equal to $L/\binom{N}{2}$.

The *local clustering coefficient* of node x is defined as the ratio of the number of links between neighbors of x to the number of all possible links between neighbors of x. Denoting the degree of node x by k_x, then if there are a_x links that connect a neighbor of x to another neighbor of x, then the clustering coefficient of node x is defined as $a_x/\binom{k_x}{2}$. In other words, clustering coefficient of node x is the number of triangles that pass through this node, relative to the maximum number of possible triangles that could pass through the node. It is more conventional to use the clustering coefficient as a network measure rather than a property of individual nodes. The clustering coefficient of a network is defined as $3N_\Delta/\sum_x \binom{k_x}{2}$, where N_Δ denotes number of triangles in the graph and the denominator is the maximum number of possible triangles. Figure 1–1 illustrates an example graph with average clustering coefficient of 3/8. Clustering coefficient is a measure of transitivity, that is, to what extent the friends of an individual are friends with one another.

A *walk* is a sequence of adjacent nodes. The sequence starts from a node and from the second node on, each node in the sequence is a neighbor of the previous node. For directed graphs, each node is the out-neighbor is the previous one. A *path* is a walk without repetition. There can be more than one paths that begin from node x and end at node y. The *graph distance* between nodes x and y is the length of the shortest path between them. If we take all the $\binom{N}{2}$ possible pairs in the network and take the average

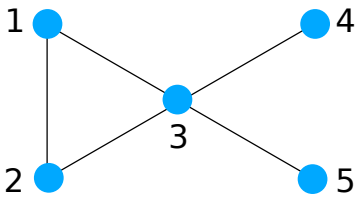


Figure 1–1: Example graph to illustrate how clustering coefficient is calculated.

of their graph distance, the result is called the *average path length* of the network. It is also sometimes called the *characteristic path length*. The *diameter* of a network is the length of the maximum graph distance between any pair of nodes.

The term 'network' was increasingly used throughout the second half of the 20th century, sometimes interchangeably with graphs. The distinction between graphs and networks is semantic and not very consequential. But just for the sake of clarity, we briefly remark on it. Graphs are mathematical representations of networks. A triangle graph, for example, can represent two distinct networks with unrelated origins: a network of three neurons forming a circle and a friendship network of three people who are friends with one another.

1.3 Historical Background on Network Science

Graph theory was first invented by Leonard Euler in his famous 'Königsberg Bridge Problem". Two islands were connected to each other and to the banks of a river. The problem was to find a walk that begins at land, passes through each bridge once, and return to the initial point. He proved this walk does not exist by mapping the lands to nodes and the bridges to links. So he invented graphs as a tool. Later, graphs were independently discovered several times. For example, Kirchhoff discovered properties of tree graphs during his work on electric circuits, Cayley worked on trees as part of enumerating organic chemical isomers, and Hamilton introduced his 'Icosian game' which led to the introduction of Hamilton Cycles. Graphs are versatile tools for modeling many systems, and here we briefly review how graph theory led to the development of modern network science and social network analysis.

The first formal network modeling of social dynamics dates back to the 1930s in the work of Jacob Moreno [Mor53]. He worked in the tradition of 'gestalt' psychology,

which is based on the idea of emergent properties irreducible to individual elements (the famous sentence 'the whole is other than the sum of its parts', sometimes wrongly translated as 'the whole is *greater* than the sum of its parts, is a quote from gestalt psychologist Kurt Koffka [Hei13]). Moreno studied how group interaction patterns limit and drive individual behavior. His primary motivation was to study how 'social aggregates' (e.g., groups, communities, cities, countries) and their 'social configurations' affect the psychology of the individual. He founded 'sociometric analysis', and founded the journal 'sociometry'. Most importantly, he was the first to draw a 'sociogram', in which individuals were connected by lines representing their relations. Although network thinking did exist as a concept in sociology (such as Simmel's 'webs of affiliation' [Sim55]), it was Moreno who initiated network thinking in its modern form.

Soon after Moreno, Fritz Heider was the next prominent figure to use sociometric tools to analyze group dynamics [Hei13]. He was interested in 'social balance', that is, how personal positive and negative attitudes translates into stable or unstable group structures. His depictions are what we now call 'signed graphs'; they consist of individuals with positive or negative links between them. These works inspired the first connection between mathematical graph theory and sociology. The first formal connection to mathematical graph theory was made by Cartwright and Harary [CH56]. Frank Harary was a mathematician who also worked on sociological problems at the time, and he made significant contributions to the development of modern graph theory. Anatol Rapoport extended the mathematical formulation to random graphs, stressing that for many biological and social systems a random-graph treatment would be more suitable [SR51, Rap57, RH61]. Around the same time, Erdős and Rényi introduced a pioneering model of random graphs. By the 1960s, graph theory had "become fashionable to mention that there are applications of graph theory to some areas of physics, chemistry, communication science, computer technology, electrical and civil engineering, architecture, operational research, genetics, psychology, sociology, economics, anthropology, and linguistics" [Har].

Stanley Milgram's famous 'small world' experiment [TM69] introduced the 'six degrees of separation' phenomenon, which brought the network conception of social relations into the popular culture. He was inspired by an idea of de Sola Pool and Kochen [dSPK78] and designed an experiment to test it. He sent a number of subjects in Omaha (Nebraska) and Wichita (Kansas) letters with instructions. Each packet had

the name of a 'target' individual in Boston. The instructions were recursively defined: If the recipients knew the target individual personally, they were to send the letter directly to the target. Otherwise, they were asked to send it to a friend or acquaintance who they thought would be more likely to know the target individual personally. Although most packets were not delivered to the targets (due to subject opt-out), using those that did, Milgram calculated the average length of chains of correspondence, which was close to 6. Although he did not use the term, 'six degrees of separation' became a pop culture term.

The sudden expansion of network science was initiated in 1998 by the pioneering work of Watts and Strogatz on the small-world model [WS98]. Inspired by Milgram's observations, they sought a model that would exhibit low average path length between the nodes (observed in the Milgram experiment). They proposed a model to "interpolate between regular and random networks" [WS98]. They reported that the neural network of the worm Caenorhabditis elegans, the power grid of the western United States, and the co-starring network between film actors, all have the small-world property. Mathematically, they call a network a small world if its average path length is the same order of magnitude as an Erdős-Rényi network with equal size and average degree, but at the same time it produces high clustering (which is also widely observed in empirical networks and the Erdős-Rényi model could not capture). Since these three networks are from completely different origins, this paper led to a great momentum of follow-ups from various fields. According to Google Scholar, since the publication of their paper, the term 'small-world network' was used in over 16000 scientific papers. This is a remarkable impact of a network model, or any model.

1.4 Social Networks and Social Outcomes

In this thesis, we focus on a specific category of networks: *social* networks. The term 'social network' is sometimes used in daily conversations to refer to online social media such as Facebook. In this thesis, we use the term to refer to any network of human relations, online or offline. Examples are friendship networks, kinship, collaboration, or connection on social media.

The field of social network analysis received a large momentum from the rapid expansion of network science mentioned above. Figure 1–2 is taken from Google NGram

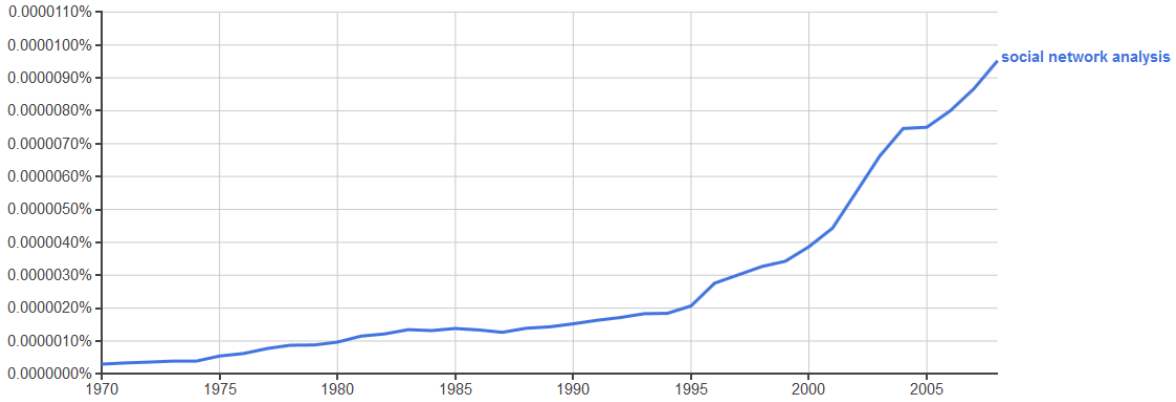


Figure 1–2: The growth in the relative occurrence of the term 'Social Network Analysis' in the corpus of the literature that Google NGram Viewer has indexed. The third word is intentionally added, because only using 'social network' would also return result that might have been related to online social media, instead of the theoretical field of research which is our purpose. Note that in the past 40 years, the relative occurrence has grown over 50 times.

viewer which searches for phrases in a large corpus of books, and returns the normalized frequency counts of their occurrences (they are normalized for each year, so an increasing trend would imply increase in the attention received in books). As can be seen, social network analysis is gathering increasing attention. Note that a factor that contributes to this growth in volume of research is the increase in data availability due to advances in technology (such as mobile phone data sets and data on online interpersonal communications).

The conceptual study of social networks and how their structure affects social outcomes is over a century old [Sim55]. Sociologists have studied how social ties that are long and weak can be conducive of novel information (because those that your strong ties provide, you probably already know) and can speed up the spread of information [Gra73] (for example, news about a job opening), how network closure and locally-dense networks can promote trust (e.g., when you have, say, 8 common friends with a person, trusting that person is less risky than if you had, say, one common friend) [Col88], and how being the only node (the *structural hole*) that connects two separate groups or communities can be advantageous in many ways [Bur09]. Social network studies have since broadened and increasingly relied on empirical data to investigate the effects of networks on social outcomes. Here we provide a few examples.

The structure of social networks also has a central role in determining public health outcomes in case of spreading diseases. Studying the effects of network topology on the spread of the disease is less than two decades old [PSV01b, PSV01a, MPSV02], and by now has been extensively studied [PSCVMV15]. Network epidemiology offers valuable insights about how the spreading behavior of a disease can be radically different for different structures of the web of contacts [LHNB15, LMVdDW11, PSCVMV15, GMT05]. Another important problem is the disease-awareness interplay [WAW+15, GGA13, FCF17] (when the prevalence of a disease increases, information and awareness spreads over the social network and people begin taking safety measures, which reduces the prevalence of the disease, which in turn lowers the panic and the safety measures, and facilitates subsequent spread of the disease). Other disease-behavior interactions, such as vaccination decisions, have also been studied in network epidemiology [NMLB+12, Bau05, BE04]. Other examples of important problems analyzed in the domain of network epidemiology include identifying the source of an epidemic (the patient-zero problem) [AFLŠ+15, ABD+14], the possible effect of infections on the topology of the underlying social network [MNHD+10], and as we will discuss below, devising optimal vaccination strategies to minimize the spread of a network [PSV02, PSV05, SMC+13, CHBA03, MKC+04].

The insight that studying the structure of social networks offers into the diffusion of information and behavior creates strong motivations for practical purposes. Identifying nodes that are more influential in the diffusion of ideas and adoptions is important in marketing [GHLH09, DJBJ10, BCZ10]. Some studies focus on the micro dynamics of how people communicate and recommend new innovations to their alters [LAH07, GGLNT04, RD02], and some studies propose methods and algorithms for finding the initial set of adopters who would maximize the diffusion (e.g., spread of adoption of a new product) over the network [KKT03, CWY09, DR01].

1.5 Thesis Outline

In Chapter 2, we present an overview of various sampling mechanisms that exist in the literature for different applications. Then we focus on sampling social networks and discuss its particular challenges and problems. We introduce a conventional social network sampling method, called the Fixed-choice Design. We argue that although it

has been used widely in various sampling designs, there is a paucity of literature on inferring the network structure from sampled data, and that most studies use the crude version of the network. In Chapter 3, we take a step towards filling this methodological gap by proposing a statistical inference framework and estimators for several network statistics.

In Chapter 5, we introduce applications where it is impractical to know the global network structure. The aim is to devise efficient sampling strategies to find influential nodes the presence of time and resource constraints. We highlight a very simple yet highly effective method that is proposed originally in the network epidemiology literature but also successfully adopted in other fields. We call it *alter sampling*. It exhibits good performance without requiring the knowledge of global network structure. Alter sampling only uses local structural information. We provide an overview of empirical studies that have successfully implemented this method in practical applications. In Chapter 6 we demonstrate that this method is highly effective across a wide array of social networks with diverse properties.

In Chapter 7, we discuss why alter sampling works so effectively. We argue that social networks must be organized and structured in certain ways in order for these local methods to work. We also highlight a related phenomenon that is observed in social networks, called the Friendship Paradox. It states that on average, people have fewer friends than their friends do. We point out that the study of the friendship paradox and its causes and consequences can help us devise effective local strategies for using structural information. Moreover, it can shed light on the global structure of social networks. We also consider the extension of the friendship paradox to personal attributes. For example, in the case of scientific collaboration networks, scholars on average have smaller H-index than their collaborators do.

In Chapter 8, we study the prevalence of the Friendship Paradox and its generalized version on online social networks. We argue that the friendship paradox and its generalizations can also be utilized to gain insight into the network structure. The results of the analysis uncovers a hierarchical nature for the connections in online social networks. We find that social networks are organized in certain hierarchical ways that in both structural and personal attributes, most people are connected to others who are on average superior. We discuss that our approach can capture structural properties that the conventional measures of correlation (such as the assortativity coefficient) fail

to grasp. We highlight the previous explanations proposed in the literature to justify this prevalence, and demonstrate that they are incomplete. That is, we show that the proposed conditions are sufficient but not necessary.

In Chapter 9, we propose a mechanistic network growth model that exhibits both the friendship paradox and its generalized version with high prevalence. We find exact mathematical expressions for how the degrees and attributes of nodes compare with those of their neighbors, which helps characterize the local inequalities.

In Chapter 10, we introduce measures to quantify the friendship paradox and the structural inequalities it relates to in the netwroks built by the proposed model in Chapter 9. These measures help characterize a network in terms of structural inequalities, so that if we have two networks, we can compare them in this term.

Finally, in Chapter 11, we discuss the consequences of the presented findings to existing results in the social networks literature, and highlight potential directions of improvement and future work.

1.6 Publications

1. N. Momeni, M. Rabbat, “Qualities and Inequalities in Online Social Networks through the Lens of the Generalized Friendship Paradox", PloS one 11.2 (2016): e0143633.
2. N. Momeni, M. Rabbat, “Measuring the Generalized Friendship Paradox in Networks with Quality-Dependent Connectivity", Complex Networks VI. Springer International Publishing, 2015. 45-55.
3. N. Momeni, M. Rabbat, “Inferring network properties from fixed-choice design with strong and weak ties", IEEE Statistical Signal Processing Workshop (SSP), 2016.
4. N. Momeni, M. Rabbat, “Generalized Friendship Paradox: An Analytical Approach", International Conference on Social Informatics. Springer International Publishing, 2014.

CHAPTER 2

Network Sampling

2.1 Chapter Outline

In this chapter we focus on network sampling, which is the first step towards any practical study on social networks. We first provide an overview of different sampling methods that are suitable in different contexts (technological networks, biological networks, online networks, etc.), and then turn our attention to the particular case of social networks. We discuss the challenges specific to sampling social networks, and introduce a conventional method called the Fixed-choice Design. We then highlight a methodological need in the literature regarding network data gathered via this sampling scheme.

2.2 Introduction

In many of the applications mentioned in Chapter 1, we need information about the global structure of the network. Basically, in any prediction task, we need to know certain measures of the global network structure. For example, consider the case of epidemic disease spreading over the network. Any theoretical prediction about the speed of contagion, prevalence as a function of time, final outbreak size, and other properties of the epidemic process during its course, depend on the structural features

of the network [PSCVMV15]. The mean and variance of the degrees, and clustering coefficient, are examples of these structural measures. In this section, we discuss the practical methods to observe the network structure and to obtain the said structural measures, discuss the limitations of these methods, and propose solutions.

The field of network sampling focuses on inferring global structural parameters from limited observed samples. There are various sampling techniques applicable to different contexts. With the rapid expansion of the Internet and other communication technologies, networked data is being produced at an increasing rate. Also, more and more massive networked data sets are being made available to researchers in various domains. Examples that have some relation to social interactions would be mobile phone data [OSH+07], online social networks [KLPM10, KNT10, WBS+09], and networks of interactions and communications on online games [ST10].

There are various sampling methods in the literature. They have been analyzed mathematically in different contexts. Here we mention a few of them. *Induced subgraph sampling* consists of selecting a random set of nodes, and collecting all the links that exist between these nodes [Kol09]. That is, a link is collected if and only if both of its incident nodes are sampled. *Incident subgraph sampling* is the converse. It first selects a set of links randomly, and then collects all the nodes that are incident to those links [Fra88, CTS13]. *Unlabeled Star Sampling* consists of first selecting a set of nodes randomly, and then collecting all the links that are adjacent to those nodes. An extension of this would be *Labeled Star Sampling*, which first takes a random sample of nodes, then collects all the adjacent links, and then collects all the nodes at the other ends of the collected links [Kol09]. Traceroute sampling [ACKM09, CM05], used in Internet applications, works as follows. A set of source nodes and a set of target nodes are selected randomly. Then, for each node in the source set, all the shortest paths to all the target nodes are specified. The sampled network consists of all the nodes and links that exist on these paths. Finally, Snowball Sampling [Kol09] is an iterated version of the labeled star sampling. We first observe an initial random set of nodes (the first wave), and observe all their links. We then collect all the nodes incident to these links (second wave), as well as all the new links incident to these nodes. Repeating the step k times would result in a k-wave snowball sampling.

In practice, most of these methods are not applicable for studying offline social networks. In the next section, we discuss limitations that are specific to offline social

networks, and the methods that are most widely used. We also highlight the severe limitations of these methods, and their consequences.

2.3 Challenges of Sampling Offline Social Networks

In sampling an offline social network, we would ideally like to sample a large number of people, and ask them to mention the name of all their friends. Then, we would find those friends, and ask them to name their friends, and so on. A sampling design in which only the first step is taken is called an *egocentric design* [Mar11, Mar90, PSC15]. One in which the latter step is also taken is called a *sociocentric design* [SCF14, PSC15]. Sociocentric design is highly time-consuming and economically prohibitive in practice. It would be possible only in very small villages (an example is [KHS+15]). Even in the rare cases that it is implemented, it has several shortcomings. Finding named alters is challenging, and sometimes impossible (for example, the named alter might have emigrated). Thus, in practice, only a small fraction of the named alters can be subsequently found and interviewed. Moreover, to increase the feasibility of finding the named alters, one will need to restrict the number of alters each respondent mentions to very few (typically one or two). This seriously reduces the breadth of the collected data.

Most social network designs are egocentric [Mar11, Mar90, MH07, PSC15]. A random sample of individuals are selected and interviewed. The respondents are asked with whom they interact. For example, they are asked with whom they have spoken in the past week, or to name their family members, or coworkers, etc. These questions are called *name generators*, because the answers to them are lists of names. Name generators specify a certain type of relationship, and ask the respondent to name alters with whom it has that type of relationship. Typically, the next step is to characterize the intensity of the relationship with each named alter. This step also involves acquiring further information about each named alter. To these ends, about each alter that the respondent mentions, a number of follow-up questions are asked. These probing questions are called *name interpreters*.

There are many practical issues in social network sampling [VT98, Mar03, PS09, KFH+10, Mar11, CL91]. Self-reported data are noisy due to significant recollection imperfections. Interview effects such as question order and satisficing are also common

practical problems. Perhaps the biggest practical problem in social network design is respondent fatigue [PS09, Joh14, Rob15]. Interviews should not be long. In long interviews, the above practical issues (recollection problems, etc.) intensify in time. Thus, the quality of the gathered data decreases in time in each interview. In the next section, we introduce the conventional way of dealing with respondent fatigue, discuss its limitations and the problems it creates.

2.4 Fixed-choice Designs

The conventional way of dealing with respondent fatigue is putting limits on the number of alters that the respondent can mention, or asking for a fixed number of alters. That is, for each name generator, the interviewer asks for a fixed number of alters, which is typically less than 10. This is widely done in practice. Here we give a few classical and a few recent examples.

For the network questions of the General Social Survey, the maximum allowable number of alters was 5 [Bur84]. A well-known classical study is by Coleman et al. on the diffusion of innovations among physicians in a small town [CKM57]. Each respondent was asked to name 3 physician friends with whom they most frequently interacted. Another pioneering study was conducted by Wellman et al. on social networks of residents of East York, Toronto. They limited the number of alters to 6 [Wel79, WCW+73]. Another famous classical social network study is on the communities in Northern California [Fis82], which limited the number of alters to 4 to 10, for different name generators. The last example is [Lau73] on friendship networks in Detroit, for which the number of allowable alters is limited to 3.

In more recent social network designs, this practice is still common. In [BKW02, BVA05, MKA+01], the number of alters are restricted to 4. In [HK07, HKC+09], the limit is 5. The studies by Christakis and Fowler on the spread of obesity [CF07], smoking [CF08], happiness [FC+08], alcohol consumption [RMFC10], loneliness [CFC09], depression [RFC11], all use the same networked data set which "captured up to two close friends" [FSC11]. Furthermore, their study of the contagion of sleep loss on social networks among adolescents limits the mentioned friends to 5 males and 5 females [MCF10]. There are many other studies that put a limit on the number of

alters [Val03, AJC64, Shu76, McC03, RSE12, CNGP07, STRM97, PSS06, BHH96, IVdBV11, VM04, BFB91, FBMG+07, Nir05, AKM+07, FMGC+07, Rei99].

A design that limits the number of alters that each respondent can name is called a *Fixed Choice Design* [WF94, New10, RH61]. It is clear that such a limit imposes an artificial cutoff on the number of personal ties and distorts the picture of the social network. Limitations of such method are pointed out by some authors [New10, WF94, SSP10], but such limitations are not mathematically discussed. Surprisingly, most studies cited above which employ the network structure for a subsequent analysis, use the crude sampled network without inference. Furthermore, no quantitative attempt for inference of the network structure from fixed-choice data exists in the literature (which might be the reason why the sampled networks are used without inference being conducted). In Chapter 3, we focus on this sampling design. We illustrate that although such a limit is imposed on the number of alters, reliable estimates on various structural quantities can be found. We theoretically find estimators for various network quantities, and verify their accuracy via numerous simulations on different topologies. The setup and results are discussed in more detail in the next chapter.

2.5 Chapter Summary

Network sampling is a necessary step for incorporating networks into any analysis. We showed that sampling offline social networks is particularly costly due to practical considerations, and it is standard to only sample a limited number of alters for each ego. For the fixed-choice design, which is a conventional method, no statistical framework for the estimation of network properties has been proposed. Chapter 3 presents our contribution to fill this gap.

CHAPTER 3

Paper: Statistical Inference for Fixed-choice Design Incorporating Strong and Weak Ties

The material presented in this chapter has been accepted in IEEE Transactions on Signal and Information Processing over Networks (the online version is available via doi: 10.1109/TSIPN.2017.2731053).

A summary of the findings is published in the following proceedings:

N. Momeni, M. Rabbat, "Inferring network properties from fixed-choice design with strong and weak ties", IEEE Statistical Signal Processing Workshop (SSP), 2016.

Please note that the references of the manuscript are listed at the end of this chapter, not at the end of the dissertation.

Inferring Structural Characteristics of Networks with Strong and Weak Ties from Fixed-Choice Surveys

Naghmeh Momeni and Michael G. Rabbat

Abstract

Knowing the structure of an offline social network facilitates a variety of analyses, including studying the rate at which infectious diseases may spread and identifying a subset of actors to immunize in order to reduce, as much as possible, the rate of spread. Offline social network topologies are typically estimated by surveying actors and asking them to list their neighbours. While identifying close friends and family (i.e., strong ties) can typically be done reliably, listing all of one's acquaintances (i.e., weak ties) is subject to error due to respondent fatigue. This issue is commonly circumvented through the use of so-called "fixed choice" surveys where respondents are asked to name a fixed, small number of their weak ties (e.g., two or ten). Of course, the resulting *crude* observed network will omit many ties, and using this crude network to infer properties of the network, such as its degree distribution or clustering coefficient, will lead to biased estimates. This paper develops estimators, based on the method of moments, for a number of network characteristics including those related to the first and second moments of the degree distribution as well as the network size, using fixed-choice survey data. Experiments with simulated data illustrate that the proposed estimators perform well across a variety of network topologies and measurement scenarios, and the resulting estimates are significantly more accurate than those obtained directly using the crude observed network, which are commonly used in the literature. We also describe a variation of the Jackknife procedure that can be used to obtain an estimates of the estimator variance.

Index Terms

Network sampling, social networks, statistical inference.

I. Introduction

A. Network Sampling

Network science has quickly spread into diverse disciplines because it offers versatile and powerful tools to quantify the structure of interactions and connections. For social networks, for instance, the diffusion of information [1]–[3] and infectious disease [4], awareness [5], and health behaviors [6], [7] are studied. The structural properties of the underlying social networks are central in these studies. Thus we need to observe and measure these properties. Like most large-scale systems, for practical considerations we need to find efficient ways of inferring these properties from a limited set of observations. This task is the focus of the network inference literature. Different sampling methods in the literature are suited for different practical requirements [8]. Examples include: traceroute sampling [9]–[11], which is typically used for sampling the Internet; respondent-driven sampling methods [12], which are typically used for sampling social networks connecting hidden populations that are difficult to find and interview; crawling methods, other random-walk methods [13], and forest fire sampling [14], which are typically used for the web and online social networks; and random node and link sampling [15], [16].

In this paper we focus on sampling offline social networks. We consider two features that are specific to social network research and that demand special consideration for network sampling. The first one involves degree truncation introduced in the measurement process, which we discuss more in Section I-B. The second one involves heterogeneity of link weights, which we discuss more in Section I-C. After introducing these two features and pointing out the absence of theoretical results on inference methods for offline social networks, we focus on incorporating them into the mathematical treatment of the sampling procedure. We introduce a setup to incorporate both of these features. We then focus on the problem of inference, which is the main contribution of this paper.

B. Fixed Choice Design

Most of the sampling methods for social networks can be mathematically formulated as variants of snowball sampling. Snowball sampling consists of sampling an initial set of

nodes and their incident links, then sampling their neighbors and their incident links, and so on. It is equivalent to running a breadth first search from the initial set of nodes, and is typically stopped at a given depth, so that not all links are traversed. Ideally, the sampling would proceed until new nodes and links are no longer encountered, so the entire network is sampled. This is impractical in most settings, and as we will discuss, even more so in offline social networks.

In practice, information about offline social networks are typically obtained through personal interviews and surveys. In this context, each person is referred to as an *ego*, and their 1-hop neighbors in the graph are called *alters*. A zero-wave snowball sample would consist of simply selecting a set of interviewees and asking them to list their alters. This is called an *ego-centric design*. For practical considerations of time and cost, the majority of social network data is ego-centric [17]–[19]. Even this simple and economical design introduces challenges, such as imperfect recollections and other memory issues. A serious practical problem is respondent fatigue, which imposes limits on the interview time and the amount of information expected from respondents. The conventional way of approaching this problem is to employ the so-called *fixed-choice design*, which amounts to imposing limits on the number of alters that each respondent is asked to list. There are numerous examples of classic and recent social networks studies that employ a fixed-choice design [20]–[25].

Interestingly, the social network studies that focus on diffusion of information, awareness, innovation and health behaviors directly use the crude, degree-truncated version of the network as the topology on which the diffusion processes take place. As pointed out recently in [26], the behavior of diffusion processes on the original networks and their degree-truncated variants can differ significantly. Thus, inferring the properties of the original network form the sampled data constitutes a significant step towards improving the results in the social network literature. The only prior works in the literature on inference of network properties from fixed-choice survey data do not differentiate between different types of social ties [27], [28].

C. Strong and Weak Ties

The second property of social networks that a sampling procedure should take into account is the heterogeneity of link weights. In social network studies, a conventional simplification is to divide social ties into strong and weak, and different questions in a survey specifically aim to elucidate different types of ties. For example, some survey questions target within-household and intimate relationships (such as secret sharing and intimate advice seeking), and others questions target between-household and weaker relations (conversations, interactions, etc.) [19], [29]. Or in the context of student friendship networks, some ties pertain to within-school friendship bonds and some pertain to between-school ties [30]. Dividing the ties into two distinct categories is a first step towards a closer correspondence to actual survey data.

Strong and weak ties have different levels of impact in different phenomena, such as diffusion of information, providing social support, adopting health behaviors, and cooperation and trust [31]–[34]. The distinct role of strong and weak social ties has been also studied in online social networks [35].

The recent study [27] discusses modeling and inference for fixed-choice designs in the simplified scenario where there is only one type of tie. In actual studies, the nature of links are heterogeneous and the first approximation would be to dichotomize them. Moreover, typical surveys used to infer structural properties of offline social networks incorporate multiple questions, while the method of [27] effectively assumes that only one fixed-choice question has been posed. While accounting for different types of ties (e.g., strong and weak) is a significant step towards making the approach more useful to sociologists, it also gives rise to a number of challenges. In particular, the model with strong and weak ties has more parameters to be estimated and requires carefully accounting for the interactions (e.g., correlations) between these parameters. Developing an inference methodology to address these challenges constitutes one of the main contributions of this manuscript, as discussed next.

D. Contribution and Paper Organization

In this paper, we study the problem of inferring network characteristics from surveys employing fixed-choice design questions. We focus on the case of networks with two distinct types of links (strong and weak). We propose an inference method to estimate network properties based on observing the sampled version of it, and we also describe a method to estimate the variance of the proposed estimators.

The rest of this manuscript is structured as follows. Section II presents the sampling setting, taking into account both features discussed above. Section III formulates the inference problem and presents methods for estimating structural properties of the network from fixed-choice survey data. Then Section IV illustrates the performance of the proposed inference methodology via simulations, and compares the results with those of the crude version of the network (without accounting for the bias introduced by fixed-choice observations).

II. Problem Formulation

Consider the following sampling setup. The *original network*, whose properties we want to estimate, is denoted by $\mathcal{G}$. This original network has N nodes, where N is an unknown parameter to be estimated. The network is undirected, and links are of one or two types: weak and strong. Thus, for each node in $\mathcal{G}$ we can define two distinct degrees, pertaining to the number of its strong links and weak links.

The sampling process starts with selecting a set of respondents (referred to as *seeds*) denoted by S_0 with cardinality $|S_0| = n_0$. Each seed is asked to name all of its strong neighbors and also B of its weak neighbors, where B is a given positive integer (see, e.g., [19]). That is, we assume that the problem of imperfect recollections can be neglected for the case of strong ties. Moreover, since the number of weak ties is typically large, we assume that the imposed limit is applied only to weak ties. We also assume that B is much smaller than the smallest weak degree in the network, so that every node has at least B weak ties to name. This is reasonable since typical fixed-choice designs use values for B that are less than ten [20]–[25].

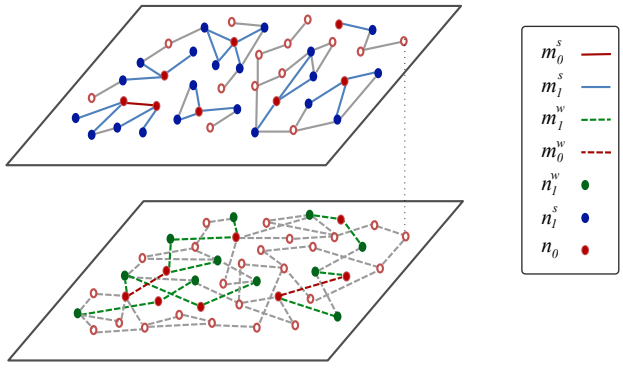


Fig. 1: A schematic illustration of the sampling setup. The upper layer represents the strong ties and the lower one represents the weak ties. The set of nodes in two layers are the same and the links in two layers are exclusive. The seeds are depicted in red. In this example $B = 2$. Gray links and hollow nodes exist in $\mathcal{G}$ but are not observed in $\mathcal{G}^*$. The observables shown in the legend are equal to the number of corresponding nodes/links (as intorduced in Section III).

The alters that each seed names might themselves belong to S_0. Let S_1^s and S_1^w denote the sets of non-seed strong and weak alters named by any seed, respectively. Note that S_0, S_1^s, and S_1^w are not disjoint, and it is possible that some node may appear in all three sets; that is, a node may be a seed, it may be named as a strong tie of another seed, and it may be named as a weak tie of yet another seed. We denote the cardinality of S_1^s by n_1^s and the cardinality of S_1^w by n_1^w. We refer to the subgraph of $\mathcal{G}$ constructed from the seeds and the responses as the *sampled network*, and we denote this sampled network by $\mathcal{G}^*$.

Figure 1 shows a schematic illustrating the sampling process. As can be seen, some of the sampled links connect two seeds, and others connect a seed to a non-seed. We denote the number of strong and weak links with both ends in S_0 by m_0^s and m_0^w, respectively. We denote the number of links between S_0 and S_1^s by m_1^s and, similarly, the number of links between S_0 and S_1^w by m_1^w. Table I provides a summary of the notation used throughout paper.

Our objective is to infer properties of $\mathcal{G}$ given the observed subgraph $\mathcal{G}^*$. We model the heterogeneity of links in the original graph with a two-layer network with the same set of

TABLE I: Notation used for statistics of the original (unknown) and sampled (observed) networks.

Original network (unknown)		**Sampled Network (observed)**	
Variable	Definition	Variable	Definition
$\mathcal{G}$	Original graph	$\mathcal{G}^*$	Sampled graph (observed)
N	Number of nodes	n_0	number of seeds
q	Sampling probablity	n_1^s	number of non-seed strong alters named by seeds
K_s	Average strong degree	n_1^w	number of non-seed weak alters named by seeds
K_w	Average weak degree	m_0^s	number of strong links between seeds
K_{ss}	Second moment of strong degrees	m_1^s	number of strong links between seeds and non-seeds
K_{ww}	Second moment of weak degrees	m_0^w	number of weak links between seeds
K_{sw}	Cross-production moment of degrees	m_1^w	number of weak links between seeds and non-seeds
T_{s^3}	Number of triangles with three strong links	$T^*_{s^3}$	Number of observed triangles with three strong links
T_{s^2w}	Number of triangles with two strong and one weak links	$T^*_{s^2w}$	Number of observed triangles with two strong and one weak links
T_{sw^2}	Number of triangles with one strong and two weak links	$T^*_{sw^2}$	Number of observed triangles with one strong and two weak links
T_{w^3}	Number of triangles with three weak links	$T^*_{w^3}$	Number of observed triangles with three weak links
τ_{ss}	Number of total triads with two strong links	λ^*_{ss}	Number of observed open triads with two strong links
τ_{sw}	Number of total triads with one strong and one weak link	λ^*_{sw}	Number of observed open triads with one strong and one weak link
τ_{ww}	Number of total triads with two weak links	λ^*_{ww}	Number of observed open triads with two weak links
λ_{ss}	Number of open triads with two strong links		
λ_{sw}	Number of open triads with one strong and one weak link		
λ_{ww}	Number of open triads with two weak links		
CC	Clustering coefficient		

nodes and two binary-valued adjacency matrices $A^s = [a^s_{ij}]$ and $A^w = [a^w_{ij}]$ representing strong and weak links, respectively. We denote the strong degree of node i by $k^s_i = \sum_{j=1}^N a^s_{ij}$ and its weak degree by $k^w_i = \sum_{j=1}^N a^w_{ij}$.

Specifically, the parameters to be estimated are the number of nodes in the original network N, the average strong and weak degrees $K_s = \frac{1}{n}\sum_{i=1}^N k^s_i$ and $K_w = \frac{1}{n}\sum_{i=1}^N k^w_i$, the second moments of the degrees (or equivalently, the variance of the degree distributions and the correlation between strong and weak degrees) $K_{ss} = \frac{1}{n}\sum_{i=1}^N (k^s_i)^2$, $K_{ww} = \frac{1}{n}\sum_{i=1}^N (k^w_i)^2$, $K_{sw} = \frac{1}{n}\sum_{i=1}^N k^s_i k^w_i$, as well as the number of triads and triangles of different types (see Sec. III-B), and the clustering coefficient [36].

III. Inference Methodology

We use the method of moments to perform inference. We need a generative model for the observables so that their expected values can be written as a function of the desired

variables. Then the method of moments proceeds by finding the least squares fit between the observables and their expected values.

We model the selection of seeds as an i.i.d. Bernoulli process, in which each node in the network is chosen as a seed independently with probability q, which is unknown. Since we seek a non-parametric framework, we also assume that the weak neighbors named by each seed are chosen uniformly at random from all of the weak neighbors of the seed. Let X_i be a Bernoulli random variable with probability q associated with each node $i = 1, \ldots, N$. If node i is a seed (i.e., $i \in S_0$ is surveyed) then $X_i = 1$, and otherwise $X_i = 0$.

Extensions to the more general case where weak neighbors are not chosen uniformly at random, but rather are chosen according to some other distribution (e.g., proportional to the neighbor's degree) may be of interest, but we leave this to future work. Likewise, it may be of interest to relax the assumption that seeds accurately report all of their strong ties (e.g., to account for forgetting one or two). Indeed, if strong ties are inadvertently omitted, then the estimates produced by the procedure described below will be biased, since they don't account for this source of error. In practice, it would be impractical to assume statistics about the number of strong ties omitted, and it would also need to be estimated. We also leave this extension to future work.

With this model and notation, our next step is to find expressions for the expected values of observed statistics n_0, m_0^s, m_1^s, n_1^s, m_0^w, m_1^w, n_1^w. Our approach to inferring the desired parameters will proceed in two stages, which we describe below. The first stage only involves estimating the first moments of the node degrees, and the second stage involves estimating the second moments.

A. First Moments

The number of nodes that are seeds can be written as $n_0 = \sum_{i=1}^{N} X_i$ and therefore we have

$$\mathbb{E}[n_0] = Nq. \tag{1}$$

Similarly, m_0^s and m_1^s can be written as $m_0^s = \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N X_i X_j a_{ij}^s$ and $m_1^s = \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N X_i(1-X_j)a_{ij}^s$, respectively. Therefore we have

$$\mathbb{E}[m_0^s] = \frac{1}{2}q^2 \sum_{i=1}^N k_i^s = \frac{1}{2}q^2 N K_s \tag{2}$$

and

$$\mathbb{E}[m_1^s] = q(1-q)\sum_{i=1}^N k_i^s = q(1-q)NK_s, \tag{3}$$

where $K_s = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N a_{ij}^s$ is the (unknown) average strong degree.

Let M_i^s be a binary variable equal to 1 if and only if node i is named as a strong neighbor by at least one seed. The total number of nodes in S_1^s is equal to $\sum_{i=1}^N (1-X_i)M_i^s$. By approximating the strong degree $\sum_{j=1}^N a_{ij}^s$ of node i by K_s, we have

$$\mathbb{E}[n_1^s] = \sum_{i=1}^N (1-q)(1-(1-q)^{k_i^s}) \simeq N(1-q)\big(1-(1-q)^{K_s}\big). \tag{4}$$

We discuss when this assumption is reasonable and investigate its consequences further in Section III-D below.

Note that if two seeds are connected with a strong link, each of them names the other one as an alter. If they are connected with a weak link, the two events corresponding to each one naming the other are assumed to be independent. We model the event that node i names node j as a weak neighbor as a Bernoulli variable W_{ij} that is equal to 1 with probability $\frac{B}{k_i^w}$ and 0 otherwise. So the total number of weak links connecting any two seeds is equal to $\frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N X_i X_j a_{ij}^w (W_{ij} + W_{ji} - W_{ij}W_{ji})$, and its expected valued can be approximated by

$$\begin{aligned}
\mathbb{E}[m_0^w] &= \frac{1}{2}\sum_{i=1}^N \sum_{j\in\mathcal{N}_i^w} q^2 \left(\frac{B}{k_i^w} + \frac{B}{k_j^w} - \frac{B^2}{k_i^w k_j^w}\right) \\
&= \frac{1}{2}q^2 B\left(2N - B\sum_{i=1}^N \sum_{j\in\mathcal{N}_i^w} \frac{1}{k_i^w k_j^w}\right) \\
&\simeq \frac{1}{2}q^2 BN\left(2 - \frac{B}{K_w}\right),
\end{aligned} \tag{5}$$

where $\mathcal{N}_i^w$ denotes the set of weak neighbors of node i. Similarly, $m_1^w = \sum_{i=1}^N \sum_{j=1}^N X_i(1-X_j)a_{ij}^w W_{ij}$ and

$$\mathbb{E}[m_1^w] = q(1-q)NB. \tag{6}$$

If we write down the expected value of n_1^w, the second moment of the degrees in the weak layer (K_{ww}) appears. As we will discuss below, K_{ww} can be estimated along with the other second moments by studying the number of triangles and triads in the observed graph. Therefore, at this step of inference it is reasonable to disregard n_1^w from the analysis.

We have six non-linear equations and four unknowns, N, q, K_w, K_s. There are different possible ways to approach solving this system of equations. One is using the generalized method of moments which minimizes the weighted squared errors of all six equations. This is infeasible because the three equations, (1), (5), and (6), admit a closed-form solution; the errors of these equations become zero, leading to the divergence of their corresponding weights. To avoid such divergence, we proceed as follows. First, using Equations (1), (5), and (6), we solve directly for $\widehat{N}$, $\widehat{q}$, and $\widehat{K}_w$:

$$\widehat{N} = \frac{Bn_0^2}{Bn_0 - m_1^w} \tag{7}$$

$$\widehat{q} = \frac{Bn_0 - m_1^w}{Bn_0} \tag{8}$$

$$\widehat{K}_w = \frac{B(Bn_0 - m_1^w)}{2Bn_0 - 2m_1^w - m_0^w}. \tag{9}$$

Then, we substitute the estimated values into (2), (3), and (4) and estimate K_s by solving the least squares problem,

$$\min_{K_s} \left[(m_0^s - \mathbb{E}[m_0^s])^2 + (m_1^s - \mathbb{E}[m_1^s])^2 + (n_1^s - \mathbb{E}[n_1^s])^2 \right]. \tag{10}$$

where the three expectations are replaced with the expressions from (2), (3), and (4).

B. Second Moments

Next we proceed to estimate the second moments of the degrees. Note that the existence of the moments of the degree distribution is not an issue here. Networks of interest in this work have finite degree moments. Diverging moments occur in heavy-tailed degree

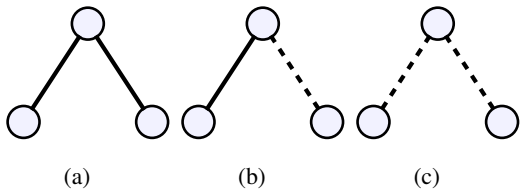


Fig. 2: Different compositions of open triads. Solid lines denote strong ties, and dashed lines denote weak ties.

distributions (e.g., power-law) *only* for infinite network size. Moreover, as mentioned above, degree distributions of offline social networks are generally much less skewed than online social networks (see [37], for example), since humans typically have limited time and capacity to maintain strong and weak ties. Thus, for the networks of interest in this work we can safely assume that the degree moments exist and are finite.

In the following, we use the term *triad* to refer to a three-node motif consisting of one node (the ego) and two of its neighbors. The neighbors can be connected (a *closed* triad) or not (an *open* triad). For example, a triangle in the original network $\mathcal{G}$ comprises three closed triads since any of the three nodes can be selected as the ego.

To employ the method of moments for estimating K_{ss}, K_{ww}, and K_{sw} and the clustering coefficient, we should again find variables that can be written as a function of the desired quantities (here, the second moments). The variables in $\mathcal{G}$ that can be written as a function of second moments are the number of different types of triads (closed and open). Due to link heterogeneity, we can have triads with different compositions, as illustrated in Figure 2.

Let τ_{ss}, τ_{sw}, and τ_{ww} denote the total number of triads in the original network, $\mathcal{G}$, similar

to the ones in Figures 2a, 2b, and 2c , respectively. Then

$$\tau_{ss} = \sum_{i=1}^{N} \binom{k_i^s}{2} \simeq \frac{1}{2} N(K_{ss} - K_s) \tag{11}$$

$$\tau_{sw} = \sum_{i=1}^{N} \binom{k_i^s}{1} \binom{k_i^w}{1} \simeq N(K_{sw}) \tag{12}$$

$$\tau_{ww} = \sum_{i=1}^{N} \binom{k_i^w}{2} \simeq \frac{1}{2} N(K_{ww} - K_w). \tag{13}$$

Note that these triads can be closed or open; the link (present or absent) between the two non-ego nodes is not accounted for here.

There are four compositions of triangles, as illustrated in Figure 3, based on the type of each link. The number of each of these triangles in $\mathcal{G}$ is denoted by $T_{s^3}, T_{s^2w}, T_{sw^2}$,and T_{w^3}. Recall that each triangle comprises three closed triads. For instance, a triangle with three strong edges gets counted as three ss triads, and a triangle with one strong edge and two weak edges corresponds to one ww triad and two sw triads.

The total number of possible triads in $\mathcal{G}$ can be written as a function of the number of open triads and the number of triangles in the network:

$$\tau_{ss} = \lambda_{ss} + 3T_{s^3} + T_{s^2w} \tag{14}$$

$$\tau_{sw} = \lambda_{sw} + 2T_{s^2w} + 2T_{sw^2} \tag{15}$$

$$\tau_{ww} = \lambda_{ww} + 3T_{w^3} + T_{sw^2}, \tag{16}$$

where $\lambda_{ss}, \lambda_{sw}$, and λ_{ww} denote the number of different open triads (Figure 2) in $\mathcal{G}$.

So based on Equations (14), (15), and (16), in order to estimate the total number of triads in $\mathcal{G}$, we need to separately estimate the number of triangles, as well as the number of open triads. To this end, we need to find the expected values of the number of triangles and open triads in $\mathcal{G}^*$ as a function of these values in $\mathcal{G}$ and the estimated variables in the first step of inference (that is, N, q, K_s, K_w).

Let us consider two illustrative examples. Consider the triangle shown in Figure 4a. Depending on whether each of the three nodes are selected as seeds, and if so whether

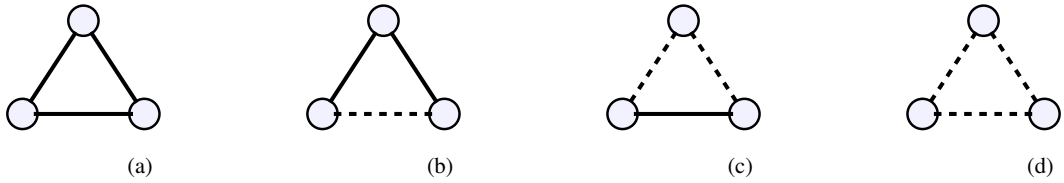


Fig. 3: Different compositions of triangles. Solid links denote strong ties, and dashed links denote weak ties.

they name the other nodes, this triangle may or may not appear in $\mathcal{G}^*$. One possible scenario in which the triangle can be observed in $\mathcal{G}^*$ is illustrated in Figure 4c. This event happens if:

1) Only nodes 1 and 2 are selected as seeds;
2) Node 1 names node 2 (and not the reverse); and
3) Node 2 names node 3 (the reverse cannot occur since node 3 is not a seed).

The probability of this event (all three points above occurring simultaneously) is $q^2(1-q)b_{11}b_{01}$, where b_{11} denotes the probability that a seed (here, node 1) names one strong link and one weak link and similarly, b_{01} denotes the probability that a seed (here, node 2) names no strong link and one weak link in the triad. These probabilities depend on the degrees of the seed. However, we approximate them for an arbitrary node x by

$$b_{01} = \frac{\binom{k_x^w-2}{B-1}}{\binom{k_x^w}{B}} = \frac{B(k_x^w - B)}{k_x^w(k_i^w - 1)} \simeq \frac{B(K_w - B)}{K_w(K_w - 1)}, \tag{17}$$

and

$$b_{11} = \frac{\binom{k_x^w-1}{B-1}}{\binom{k_x^w}{B}} = \frac{B}{k_x^w} \simeq \frac{B}{K_w}. \tag{18}$$

Similarly, we can define b_{00}, b_{10}, b_{20}, and b_{02}. Their approximated expressions are shown in Table II in the Appendix. There are 42 possible ways that a triangle in $\mathcal{G}$ can be observed in $\mathcal{G}^*$, and these factors can be used as building blocks for calculating the probabilities pertaining to all 42 possible ways. We denote the probability of observing triangles in $\mathcal{G}^*$ by $\{\rho_j; j = 1, 2, ..., 42\}$. All of the triangles and corresponding expressions for ρ_j are presented in Figure 12 and Table III in the Appendix.

(a) Triangle in $\mathcal{G}$ (b) Open triad in $\mathcal{G}$ (c) Triangle in $\mathcal{G}^*$ (d) Open triad in $\mathcal{G}^*$

Fig. 4: Example of a triangle and an open triad in $\mathcal{G}$ being observed in $\mathcal{G}^*$. The hollow nodes are chosen as respondents, and the solid node is not chosen. Solid lines represent strong links. Dashed lines represent weak links. Arrows indicate mentioning the adjacent node in the interview.

The same triangle in Figure 4a can be observed as an open triad in $\mathcal{G}^*$. One possible scenario is illustrated in Figure 4d. The probability of this event is equal to $q^2(1-q)b_{11}b_{00}$. There are 31 possible ways an open triad can be observed in $\mathcal{G}^*$ (see Figure 13). We denote the probability of observing triangles in $\mathcal{G}$ as open triads in $\mathcal{G}^*$ by $\{\pi_i; i = 1, 2, ..., 31\}$ (Table IV).

Open triads in $\mathcal{G}^*$ are not observed whenever at least one link in a triangle in $\mathcal{G}$ is absent (not named). They can be observed if an open triad with the same composition is preserved during the sampling process. Consider again the open triad in Figure 4d. It can originate from the triangle in Figure 4a or from the open triad in Figure 4b. Note that in the latter case, node 2 in not connected to node 3 in $\mathcal{G}$. So the absence of this link in $\mathcal{G}^*$ is not the result of node 2 not naming node 3 (unlike the case of triangle to triad). We can write the probability of observing this triad originating from the triad in Figure 4b as $q^2(1-q)b_{11}a_{00}$, where a_{00} corresponds to node 2 not naming any strong or weak link while it is connected to only one of them. For an arbitrary node x, a_{00} can be approximated by

$$a_{00} = \frac{\binom{k_x^w - 1}{B}}{\binom{k_x^w}{B}} = \frac{k_x^w - B}{k_x^w} \simeq 1 - \frac{B}{K_w}. \tag{19}$$

Similarly, we can define a_{01} and a_{10}. The approximation of these quantities are presented in Table II in the Appendix. We denote the probability of observing open triads in $\mathcal{G}$ as open triads in $\mathcal{G}^*$ by $\{\phi_i; i = 1, 2, ..., 31\}$ (Table V).

Let us denote the expected number of different types of triangles and open triads in $\mathcal{G}^*$ with the same notation introduced for the original network with the addition of $*$

superscrtipts. For the triangles we have

$$T^*_{s^3} = T_{s^3} \times \sum_{i=1}^{2} \rho_i \tag{20}$$

$$T^*_{s^2w} = T_{s^2w} \times \sum_{i=3}^{7} \rho_i \tag{21}$$

$$T^*_{sw^2} = T_{sw^2} \times \sum_{i=8}^{17} \rho_i \tag{22}$$

$$T^*_{w^3} = T_{w^3} \times \sum_{i=18}^{26} \rho_i. \tag{23}$$

For example, Equation (21) states that the s^2w triangles in $\mathcal{G}$ can be observed in $\mathcal{G}^*$ under 5 different events, depicted in Figure 12, whose probabilities are listed in Table III. A similar explanation and reasoning follows for the other triangular configurations. Also, the expected number of open triads in $\mathcal{G}^*$ can be written as

$$\lambda^*_{ss} = 3T_{s^3} \times \pi_3 + T_{s^2w} \times \sum_{i=1}^{4} \pi_i + \lambda_{ss} \times \sum_{i=1}^{4} \phi_i \tag{24}$$

$$\lambda^*_{sw} = 2T_{s^2w} \times \pi_6 + 2T_{sw^2} \times \sum_{i=5}^{13} \pi_i + \lambda_{ss} \times \sum_{i=5}^{13} \phi_i \tag{25}$$

$$\lambda^*_{ww} = T_{sw^2} \times \pi_{14} + 3T_{w^3} \times \sum_{i=14}^{24} \pi_i + \lambda_{ww} \times \sum_{i=14}^{24} \phi_i. \tag{26}$$

Note that the coefficients $a_{i,j}$ and $b_{i,j}$ in Table II are all functions of B and K_w. Similarly, the coefficients ρ_i, π_i, and ϕ_i only depend on the values from Table II and q. Since the parameter B is assumed to be known, given estimates of q and K_w we can approximate all of these coefficients. Then we can use the estimated values in conjunction with Equations (20)–(26) to estimate the number of triangles and triads. Note that, given the coefficients, these are all linear equations in the unknown parameters, so estimation reduces to solving a system of linear equations. Finally, we use the estimated numbers of triangles and triads to estimate the degree correlations, K_{ss}, K_{sw}, and K_{ww} via Equations (11), (12), and (13), which are also linear equations in the unknowns.

C. Summary of the Inference Method

The following steps summarize the entire proposed inference method.

1) Estimate $\widehat{N}, \widehat{q}$, and $\widehat{K}_w$ (Equations (1), (2), and (3)).
2) Estimate the strong degree $\widehat{K}_s$ (10).
3) Plug in the estimated weak degree $\widehat{K}_w$ and sampling probability $\widehat{q}$ to estimate values for $\{\rho_j; j = 1, 2, ..., 26\}$ and $\{(\pi_i, \phi_i); i = 1, 2, ..., 24\}$ (Tables III, IV, V).
4) Count the number of observed triangles ($T^*_{s^3}$, $T^*_{s^2w}$, $T^*_{sw^2} = T_{sw^2}$, and $T^*_{w^3} = T_{w^3}$) and triads (λ^*_{ss}, λ^*_{sw}, and λ^*_{ww}) in $\mathcal{G}^*$.
5) Estimate the number of different triangles in $\mathcal{G}$ (Equations (20), (21), (22), (23)).
6) Estimate number of different open triads in $\mathcal{G}$ (Equations (24), (25), and (26)).
7) Estimate the total number of all triads in $\mathcal{G}$ (Equations (14), (15), and (16)).
8) Estimate $\widehat{K}_{ss}$, $\widehat{K}_{sw}$, and $\widehat{K}_{ww}$ (Equations (11), (12), and (13)).

The computational complexity of this method is dominated by step 4, which involves counting all triangles and triads in the observed network. Typical studies of offline social networks focus on villages populations smaller than 10^4. For observed networks of this size, running the entire inference procedure takes about one second on a contemporary laptop computer.

D. The Average Degree Approximation

Many steps of the development above involve approximating the individual node degrees k^s_i and k^w_i with the average values K_s and K_w. Let us briefly describe why this is both practically and theoretically reasonable. First note that the less skewed the degree distribution is, the better the said approximation performs. Although there is no social network study in which a full real-world offline social network has been observed, there are studies which provide the degree distribution. For example, see Figure 1 in [37]. Offline social networks exhibit reasonably concentrated degree distributions, not heavy tailed. So it is expected that the adopted approximation is not a significant source of error.

Let us also estimate the error theoretically. Consider a network whose weak and strong degree distributions are both Poisson; i.e., suppose k_i^s, $i = 1, \ldots, N$ are i.i.d. Poisson random variables with mean K_s. It is straightforward to show that

$$\mathbb{E}[\sum_i (1-q)^{k_i^s}] = Ne^{-qK_s}. \tag{27}$$

Evaluating the Taylor expansion of this expression at $q = 0$, we find that the relative error is

$$\frac{\mathbb{E}[\sum_i (1-q)^{k_i^s}]}{N(1-q)^{K_s}} = 1 + \frac{K_s q^2}{2} + \frac{K_s q^3}{3} + \frac{K_s(K_s+2)q^4}{8} + O(q^5).$$

Thus, the leading term in the error is proportional to q^2, which is reasonably small (note that in typical offline social networks, K_s may be a few 10's while q is significantly smaller than one).

For Equation (5), theoretical calculation cannot be performed in closed form even for Poisson networks. Thus, to verify its applicability, we tested it on a variety of network models with different properties. Figure 5 presents the distributions of $\frac{\hat{Y}}{Y}$, where Y is the sum $\sum_i \sum_{j \in \mathcal{N}_i^w} \frac{1}{k_i^w k_j^w}$, approximated in Equation (5) by the expression $\widehat{Y} = \frac{N}{K_w}$. In the simulation we employ four distinct families of networks with different properties to investigate the robustness of the approximation. The four synthetic network models are: Small-world (SW) [38], Barabasi-Albert (BA) [39], Random Recursive Trees (RRT) [40], and the high-clustering scale-free model of Holme and Kim (HK) [41]. The difference between BA and RRT is that in the BA model incoming nodes choose their neighbors preferentially (i.e., with degree-proportional probabilities), whereas in the RRT model they choose them uniformly at random. The BA and RRT models generate networks with unrealistically high-skewed degree distributions for offline social networks; we present them here as extreme worst-case scenarios.

We generate 1000 random networks from each family, with parameters randomly generated, and with sizes fixed at $N = 1000$. The distribution of relative errors is presented in Figure 5. It can be observed that the relative error for these networks is less than 6%. For the HK and SW models, which may be considered more realistic models of offline social

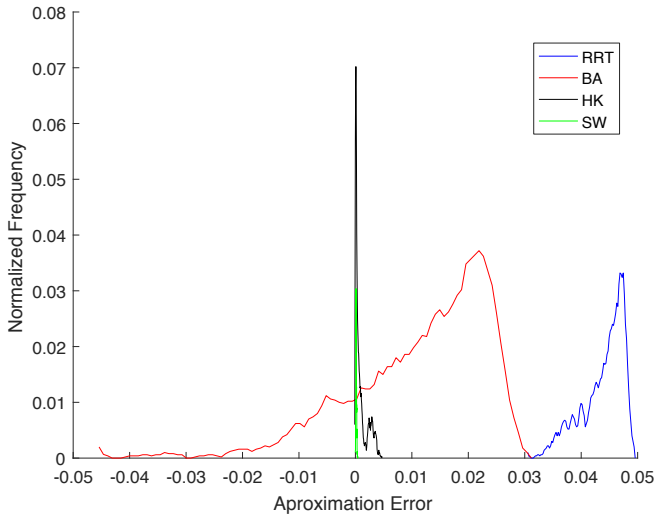


Fig. 5: The distribution of relative error due to the approximation made in Equation (5) for the SW, RRT, BA, and HK families of networks. For each family we generated 1000 networks of size 1000, with parameters randomly generated. As mentioned in the text, BA and RRT are worst-case scenarios due to their extreme skew, yet the relative error is reasonably small for them. For the more realistic models of SW and HK, the error is considerably smaller.

networks, the relative error is about 1%. This demonstrates the reasonable accuracy of the approximations.

IV. Results and Discussions

A. *Performance of the Proposed Estimators*

To verify the accuracy of the estimators, the ideal scenario would be to have data from real-world offline social networks that have been fully observed (i.e., ground truth), along with their sampled versions. We found no fully-observed real-world offline social network dataset available in the literature. Full observation is almost impossible due to practical and privacy considerations. Thus we use synthetic networks for evaluation.

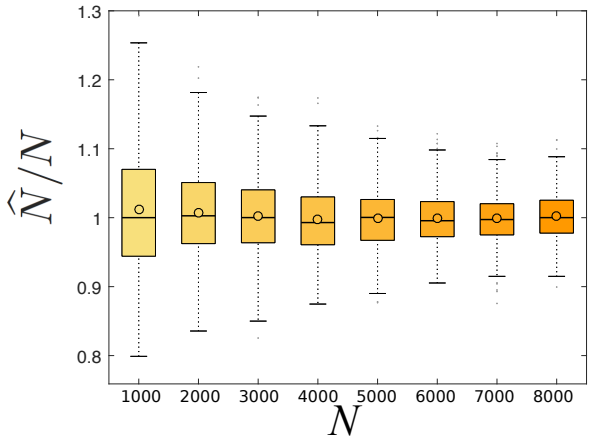
We verify the accuracy of the proposed estimators via Monte Carlo simulations over 500 synthetic networks. In each MC trial, we build a synthetic two-layer network. The set of nodes in the two layers is the same. One layer represents the strong links and the other

represents weak links. All the synthetic networks are generated according to a modified Watts-Strogatz model, with the difference being that edges are randomly added instead of being randomly rewired [38]. We randomly sample model parameters such that the average degree of the weak layer falls between 100 and 200 and the average degree of the strong layer is between 10 and 20. These values are justified by the substantial literature in evolutionary psychology and neuroscience [42]–[48] which suggests that the human brain has evolved to maintain approximately 150 active social ties, with an 'inner circle' of up to 20 members. The results are not sensitive to these precise values; increasing the average degree in the weak layer does not substantially change the results.

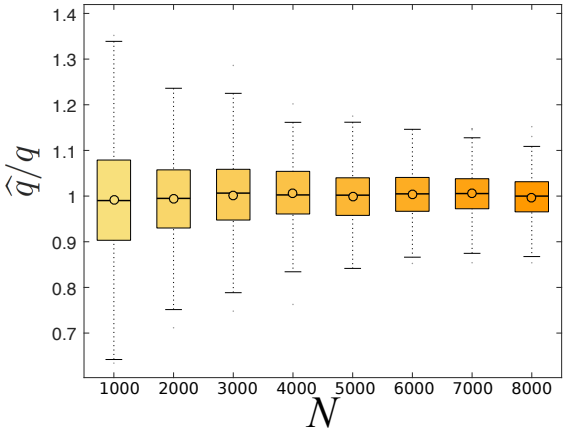
We apply the sampling process on this network. Then, we infer the desired variables and compare them to the true values. The number of weak links named by seeds is $B = 10$. We first keep $q = 0.1$ constant and increase N to confirm that the performance improves as the network size increases. We then fix $N = 4000$ and vary q to study the effect of sampling proportion. The same simulations are repeated with $B = 2$. Finally, we fix $N = 4000$ and $q = 0.1$ and vary B.

Figure 6 shows the empirical distribution of the ratio of the estimated values to the true values for N, q, K_s, and K_{ss} (all for $B = 10$). In all cases, the estimator does exhibit some bias for smaller sizes of networks, N, and samples, q. As the number of nodes or the sample size increases the variability of the estimates decreases, as does the bias. The results when $B = 2$ are similar to the case of $B = 10$ (see the supplementary material); only the variability of the estimates is greater, but not significantly so.

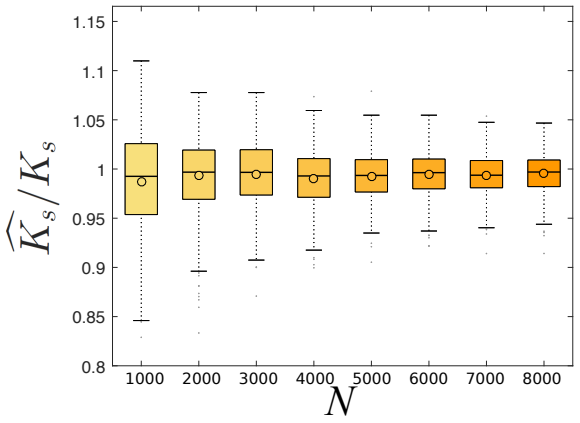
Figure 7 presents the results for K_w, K_{sw}, and K_{ww}. For these estimators, the results depend on the value of B. The results for K_{sw} and K_{ww} resemble those of K_w, so we omitted multiple figures. For $B = 10$, the estimates are not biased for different sizes of network and samples and their variability consistently decreases as the size increases (Figures 7a and 7c). However, the behaviour of the estimators is different when $B = 2$. In Figure 7b, the estimator has a bias for smaller values of N and the bias decreases as N increases. In Figure 7d, the estimator shows a significant bias for $q = 0.05$. As q increases, first the bias and then the variability of the estimator are improved. Figures 7e and 7f



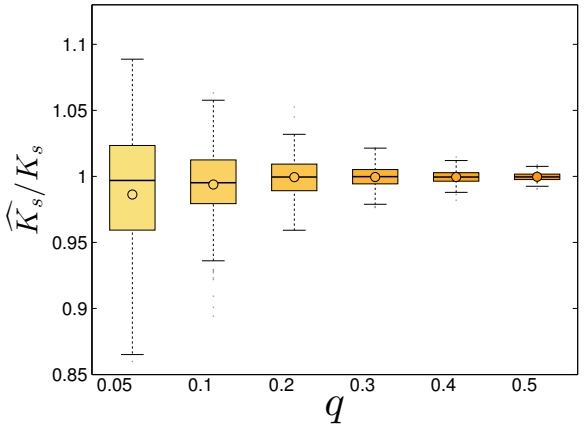
(a) Number of Nodes, $q = 0.1$, $B = 10$



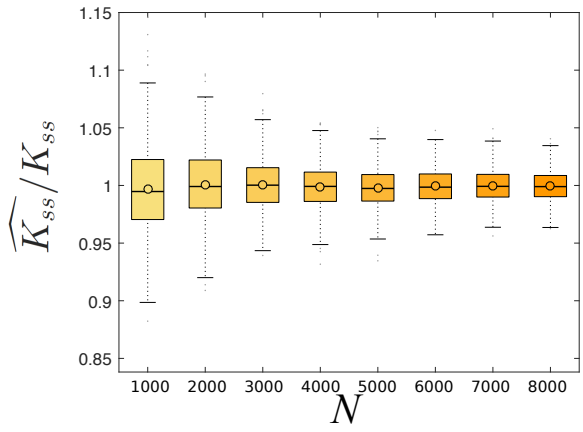
(b) Sampling probability, $q = 0.1$, $B = 10$



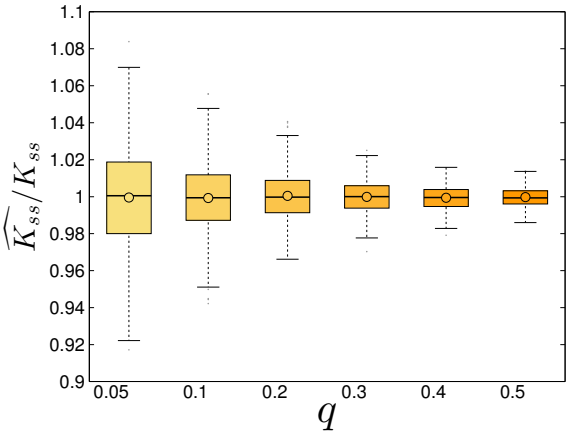
(c) Average strong degree, $q = 0.1$, $B = 10$



(d) Average strong degree, $N = 4000$, $B = 10$



(e) Second moment of strong degrees, $q = 0.1$, $B = 10$



(f) Second moment of strong degrees, $N = 4000$, $B = 10$

Fig. 6: Distribution of ratio of estimated values and true values.

illustrate the dependence of the performance of the estimators on B. It can be seen that the bias is negligible for values of B as small as 4. If we further increase B, the variability of estimates decreases.

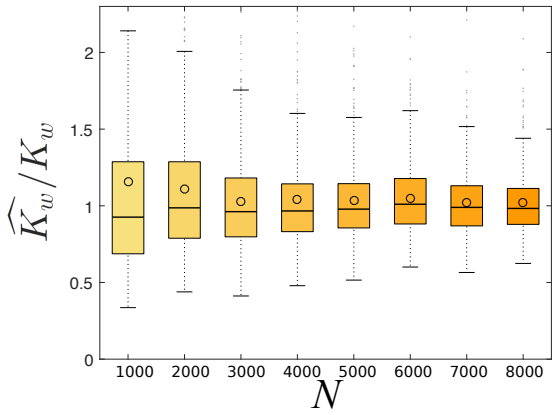
To test the robustness of the results on moderate levels of skew that might be observed in offline social networks, we also tested the results on the high-clustering scale-free model of Holme and Kim [41], and the observed results are reasonably accurate. Since the results are similar to those presented, we omit them for space limitations.

B. Comparing with SFC Estimators

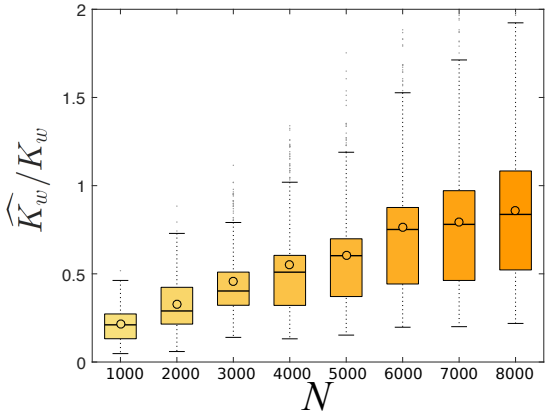
Next, we compare the performance of the method proposed in this paper to the one proposed in [27], which we refer to as the *single fixed choice* (SFC) method, since it draws inferences about network structure based on the responses to a single fixed-choice survey question without differentiating between weak and strong ties. In order to facilitate this comparison, we use the same synthetic networks as described above. We first apply the proposed sampling and inference method. Then we collapse the two-layer network into a single network and apply the SFC sampling and inference scheme. We compare the performance of the two methods in terms of their estimate of the average degree, we take the estimates for K_s and K_w produced by the proposed method and compare the $K_s + K_w$ with the estimate of K produced by SFC. Figures 8a and 8b show the performance of estimators of number of nodes and the average degrees. Although the estimates of the network size are comparable, the estimates of average degree are slightly better for the proposed approach, with it having a smaller inter-quartile range. This is not surprising, since the proposed method produces a better estimate of the average strong degree, thereby providing a more reliable estimate of the total degree.

C. Comparing Results with the Crude Version

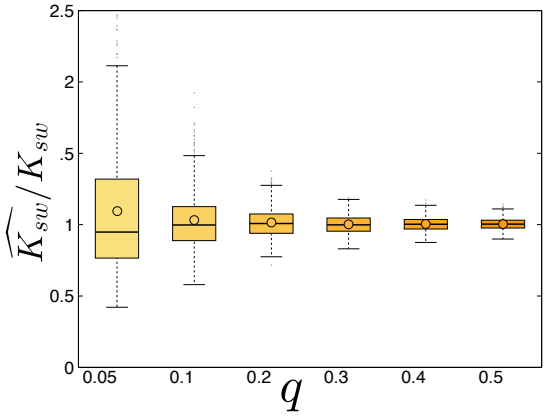
As discussed in Section I, many social network studies of contagion use the sampled network without any inference [20]–[25]. To compare our estimators with the crude values of the sampled network, we need to choose a network statistic. The effect of degree



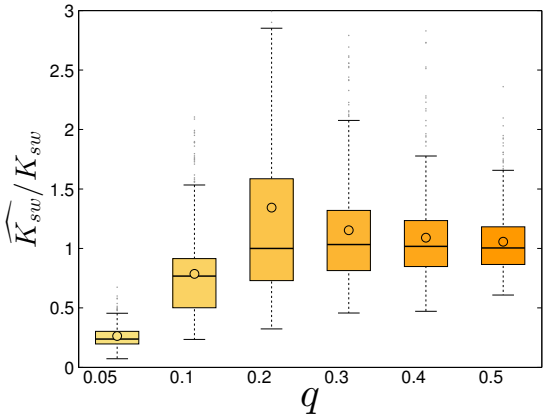
(a) Average weak degree, $q = 0.1$, $B = 10$



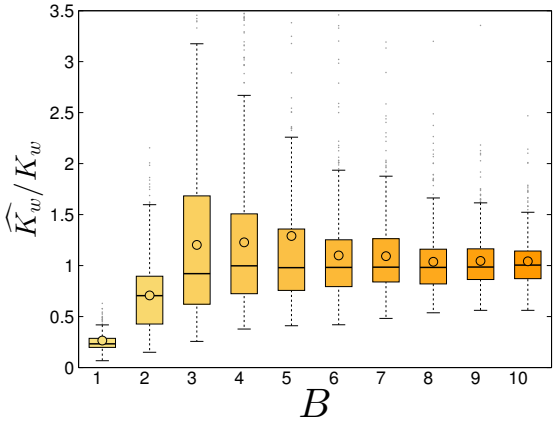
(b) Average weak degree, $q = 0.1$, $B = 2$



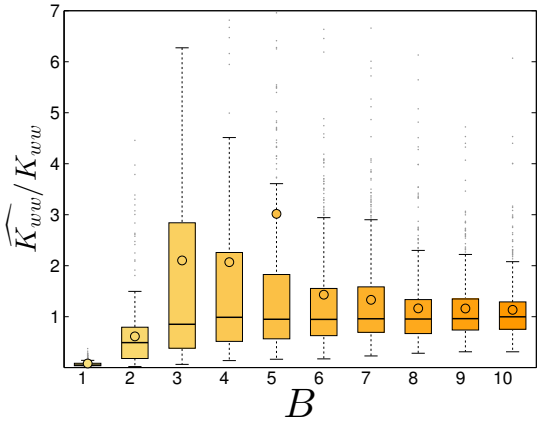
(c) degrees product moment, $N = 4000$, $B = 10$



(d) degrees product moment, $N = 4000$, $B = 2$

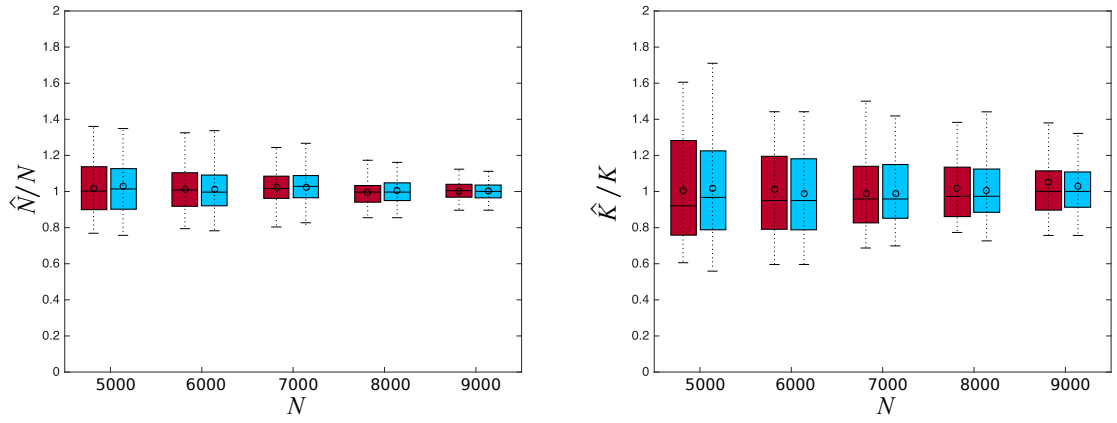


(e) Average weak degree, $N = 4000$, $q = 0.1$



(f) Second moment of weak degrees, $N = 4000$, $q = 0.1$

Fig. 7: Distribution of ratio of estimated values and true values.



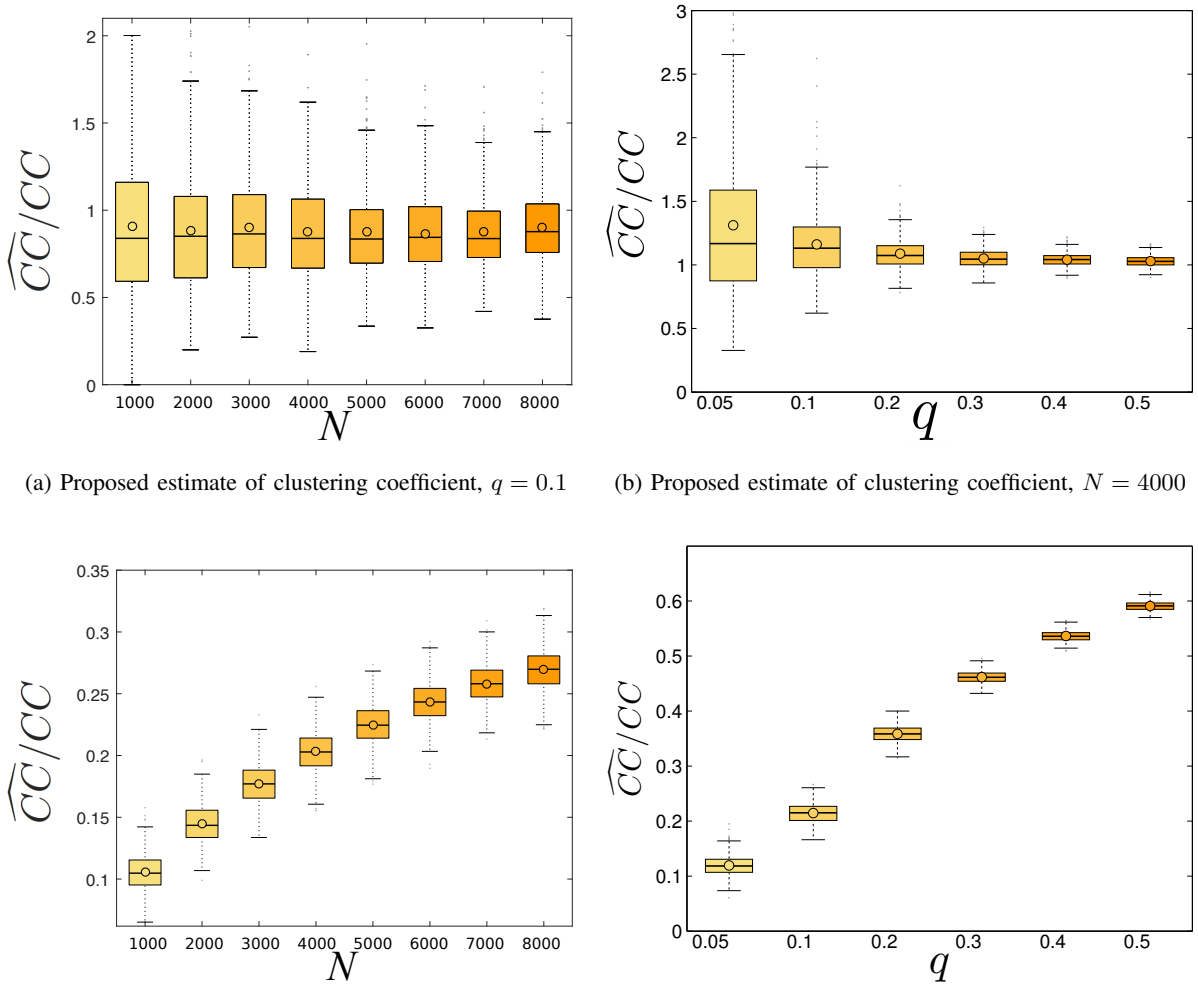
(a) Number of Nodes, $q = 0.1$, $B = 10$

(b) Average Degree, $q = 0.1$, $B = 10$

Fig. 8: Comparing the performance of our estimators to SFC method [27]. Red boxes (left, within each group) show the distribution of SFC estimates and blue boxes (right, within each group) show the distribution of our estimates. (Best viewed in color.)

truncation in moments of the degree distribution is trivial, and it is clear that the crude values will be heavily biased, in comparison to the values produced by our estimators which appear to exhibit good performance in the experiments reported above.

Instead, we consider estimating the clustering coefficient, a dimensionless quantity which is one of the most important network statistics in social network studies [49]. It is also desirable because it embodies all the other estimators and approximations. Since we have different types of triads and triangles, we first collapse the network into one layer (with homogeneous links) and then calculate its clustering coefficient. Figures 9a and 9b illustrate the performance of our method in estimating the clustering coefficient. Figures 9c and 9d depict the clustering coefficient calculated directly from the crude sampled network. It is evident that our approach outperforms the crude estimate by a large margin, and using the crude estimates results in underestimating the clustering coefficient of the network.



(a) Proposed estimate of clustering coefficient, $q = 0.1$

(b) Proposed estimate of clustering coefficient, $N = 4000$

(c) Crude estimate of clustering coefficient, $q = 0.1$

(d) Crude estimate of clustering coefficient, $N = 4000$

Fig. 9: Comaparing our estimates and crude estimates of clustering coeffcient for $B = 10$

D. Performance of Estimators on Real-world Datasets

The Villages dataset [29] consists of surveys made in 77 villages in India. The questionnaire includes several questions used to build the social networks. To apply our method to the networks in this data set, we form the strong layer of each network based on responses to a question about relationships involving the borrowing of money, and we build the weak layer by connecting two nodes with a weak tie if they are not relatives and accompany

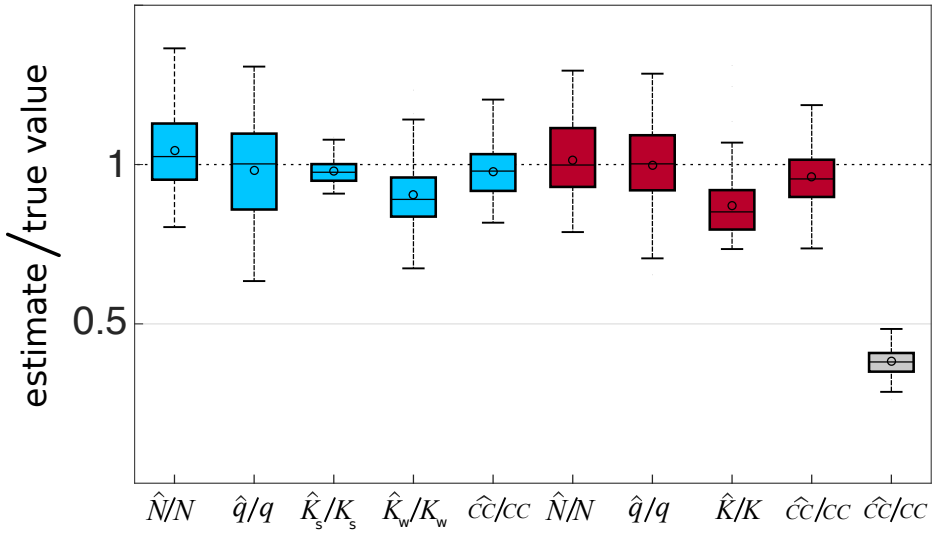


Fig. 10: Performance on the Villages dataset. Red boxes correspond to the proposed estimator, blue boxes correspond to the SFC estimator, and the grey box corresponds to the crude estimate of the clustering coefficient.

each other when going to temple. We remove the nodes whose weak degree was less than 3 and then applied the sampling method with $B = 3$. The distribution of the estimates for N, q, K_s, K_w and the clustering coefficient (for the collapsed network) using the proposed method are presented in Fig 10 in red. To compare our estimates to the SFC model, we collapse the two layers into one and estimate N, q, K, and clustering coefficient using SFC. The distribution of these estimates are shown in Fig 10 in blue. Also, we have included the crude estimates of the clustering coefficient (calculated as explained in Section IV-C) in gray. As with the simulated dataset, the clustering coefficient estimate obtained using the proposed method is significantly better than that obtained using the crude network. Moreover, the accuracy of the proposed approach in estimating a variety of network parameters provides some validation that the modeling assumptions on which this approach is based are reasonable.

V. Estimating the Variance of the Estimators

To estimate the variance of the estimators, we propose a variation of the Jackknife resampling method [50], [51]. In each resampling, we leave out one of the respondents and remove all the links of that respondent in the sampled network. Then, we apply our method to estimate the desired variables in the resampled network. Variances are estimated from the distribution that is obtained by repeating this procedure for all respondents. Note that this estimated variance is different from the variance of all estimates from subsamples. The estimated variance of an estimator for parameter h in this method is equal to

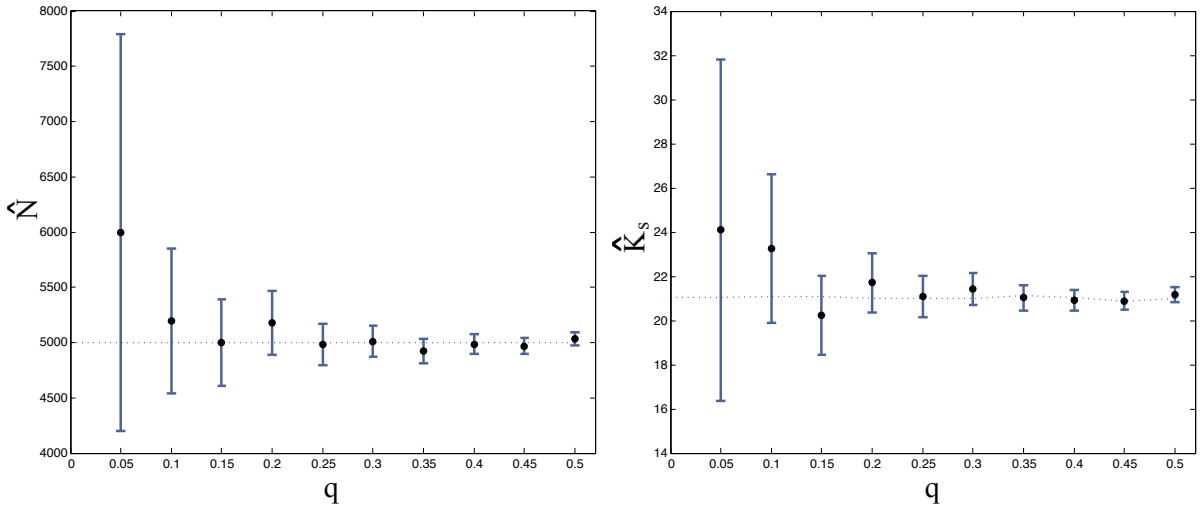
$$Var(h) = \frac{n_0 - 1}{n_0} \sum_{n=1}^{n_0} (\widetilde{h}_i - \bar{h})^2, \tag{28}$$

where $\widetilde{h}_i$ is the estimated value of h when node i is removed from the seeds and $\bar{h}$ is the average of all values of $\widetilde{h}_i$. Figure 11 presents the results of Jackknife resampling for two of the estimators for different values of sampling probability ($N = 5000$ and $B = 10$ are fixed). It can be seen that as the sample size increases, the estimated standard deviations of both estimators decrease. Moreover, for all values of q we see that the true value (dashed line) falls within one standard deviation of the jackknife-estimated mean.

VI. Conclusion and Future Work

This paper described a method for estimating characteristics of a social network topology (the network size, average number of strong and weak ties, as well as second moments of the strong and week degree distributions) from fixed choice survey data. In particular, we assumed that every respondent provides all of their strong ties and a fixed number of their weak ties. The proposed estimation methodology is based on the method of moments, under a model where respondents are sampled according to a Bernoulli process over vertices (with unknown sampling rate) and the subset of reported weak ties is sampled uniformly from all of the respondents weak ties.

A natural extension of the work proposed is to consider surveys with a soft fixed choice design; instead of reporting exactly B weak ties, each respondent may report up to B weak ties. One approach to this may be to assume that each respondent x samples a number



(a) Estimator for N

(b) Estimator for K_s

Fig. 11: Standard deviation of two estimators. The circles represent the jackknife estimates of the mean and the error bars represent the jackknife estimate of the standard deviation. The dashed line represent the true values.

$B_x \leq B$ of weak ties to report with B_x being independently and identically distributed with an unknown mass function over the integers from 0 to B. In this case, in the context of the model developed in this paper, it turns out that it is sufficient to estimate the mean $\mathbb{E}[B]$. We are currently exploring such an estimator, as well as theoretical guarantees for the proposed inference procedure. Another possibility is to make parametric assumptions about the recollection process and to modify the assumption of seeds choosing weak ties uniformly at random.

In social health-related applications, it is commonly of interest to identify a subset of the population to be immunized, with the intention of most efficiently preventing the spread of infectious diseases, subject to a constraint on the number of individuals that can be immunized. For this reason, it would also be of interest to extend the results of this paper to estimate quantities such as the betweenness centrality (or another centrality) measure of each node, since these typically correlate highly with individuals that are well-placed (i.e.,

TABLE II: Approximate probabilities corresponding to the outcome of the sampling process for seeds which are part of triangles and open triads

$b_{00} \simeq \frac{(K_w-B)(K_w-B-1)}{K_w(K_w-1)}$	$b_{01} \simeq \frac{B(K_w-B)}{K_w(K_w-1)}$	$b_{11} \simeq \frac{B}{K_w}$
$b_{02} \simeq \frac{B(B-1)}{K_w(K_w-1)}$	$b_{11} \simeq \frac{B}{K_w}$	$b_{20} = 1$
$a_{00} = \simeq 1 - \frac{B}{K_w}$	$a_{01} \simeq \frac{B}{K_w}$	$a_{10} = 1$

hubs) in the network.

APPENDIX

Here we present additional figures and tables of expressions used in the estimator calculations described in Section III. Table II summarizes the expressions for the coefficients a_{ij} and b_{ij}, $i, j \in \{0, 1\}$, used for calculating estimates of the second moments.

Figure 12 shows all of the 42 possible ways that a triangle in $\mathcal{G}$ can be observed in $\mathcal{G}^*$. Table III shows the corresponding expressions ρ_j for each $j = 1, \ldots, 26$ corresponding to each example shown in Figure 12.

Figure 13 shows the 31 possible ways that an open triad in $\mathcal{G}$ can be observed in $\mathcal{G}^*$, and Table IV provides expressions for the probability π_i, $i = 1, \ldots, 31$ of observing each one. Table V provides expressions for the probability ϕ_i, $i = 1, \ldots, 31$, of observing each open triad as as an open triad in $\mathcal{G}^*$.

REFERENCES

[1] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proc. 21st Intl. Conf. on World Wide Web*. ACM, 2012, pp. 519–528.

[2] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *IEEE 10th Intl. Conf. on Data Mining*, 2010, pp. 599–608.

[3] Y. Moreno, M. Nekovee, and A. F. Pacheco, "Dynamics of rumor spreading in complex networks," *Physical Review E*, vol. 69, no. 6, p. 066130, 2004.

[4] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Review of Modern Physics*, vol. 87, pp. 925–979, 2015.

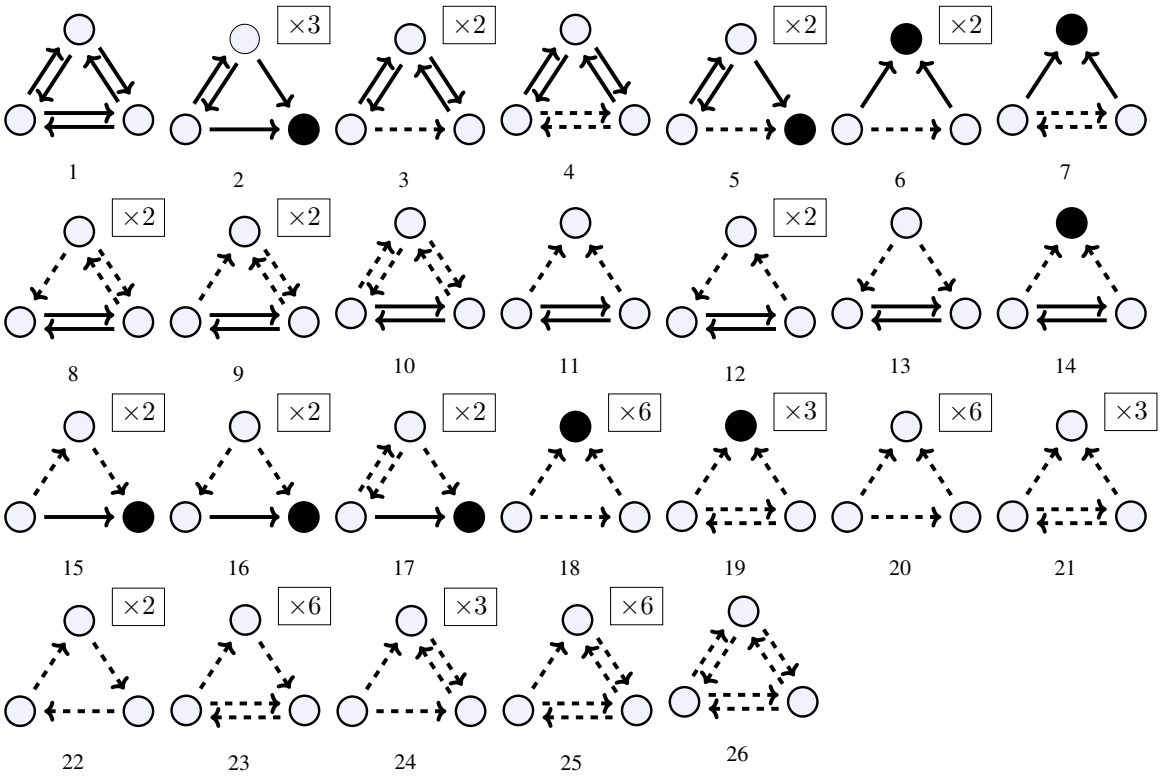


Fig. 12: All possible ways a triangle can be observed in $\mathcal{G}^*$. Some configurations are degenerate (due to symmetry). For these cases, the corresponding multiplicities are given on the top right.

TABLE III: Probabilities of observing different triangles

$\rho_1 = q^3 b_{20}^3$	$\rho_2 = 3q^2(1-q) b_{20}^2$	$\rho_3 = 2q^3 b_{20} b_{11} b_{10}$	$\rho_4 = q^3 b_{20} b_{11}^2$
$\rho_5 = 2q^2(1-q) b_{20} b_{11}$	$\rho_6 = 2q^2(1-q) b_{11} b_{10}$	$\rho_7 = q^2(1-q) b_{11}^2$	$\rho_8 = 2q^3 b_{02} b_{10} b_{11}$
$\rho_9 = 2q^3 b_{01} b_{11}^2$	$\rho_{10} = q^3 b_{02} b_{11}^2$	$\rho_{11} = q^3 b_{00} b_{11}^2$	$\rho_{12} = 2q^3 b_{01} b_{11} b_{10}$
$\rho_{13} = q^3 b_{02} b_{10}^2$	$\rho_{14} = q^2(1-q) b_{11}^2$	$\rho_{15} = 2q^2(1-q) b_{01} b_{11}$	$\rho_{16} = 2q^2(1-q) b_{02} b_{10}$
$\rho_{17} = 2q^2(1-q) b_{02} b_{11}$	$\rho_{29} = 6q^2(1-q) b_{02} b_{01}$	$\rho_{19} = 3q^2(1-q) b_{02}^2$	$\rho_{20} = 6q^3 b_{00} b_{01} b_{02}$
$\rho_{21} = 3q^3 b_{00} b_{02}^2$	$\rho_{22} = 2q^3 b_{01}^3$	$\rho_{23} = 6q^3 b_{01}^2 b_{02}$	$\rho_{24} = 3q^3 b_{01}^2 b_{02}$
$\rho_{25} = 3q^3 b_{01} b_{02}^2$	$\rho_{26} = q^3 b_{02}^3$		

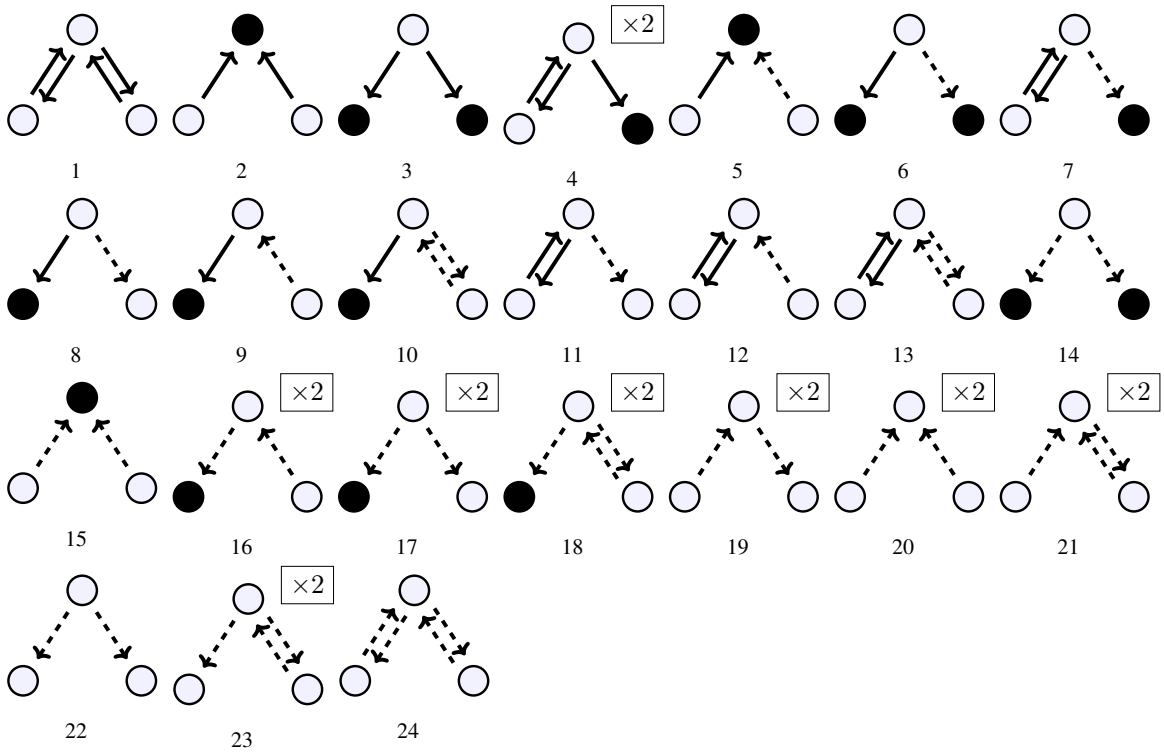


Fig. 13: All possible ways an open triad can be obserevd in $\mathcal{G}^*$.

TABLE IV: Probabilities of observing different open triads originated from triangles

$\pi_1 = q^3 b_{20} b_{10}^2$	$\pi_2 = q^2(1-q) b_{10}^2$	$\pi_3 = q(1-q)^2 b_{20}$	$\pi_4 = 2q^2(1-q) b_{20} b_{10}$
$\pi_5 = q^2(1-q) b_{10} b_{01}$	$\pi_6 = q(1-q)^2 b_{11}$	$\pi_7 = q^2(1-q) b_{11} b_{10}$	$\pi_8 = q^2(1-q) b_{11} b_{00}$
$\pi_9 = q^2(1-q) b_{10} b_{01}$	$\pi_{10} = q^2(1-q) b_{11} b_{01}$	$\pi_{11} = q^3 b_{10} b_{11} b_{00}$	$\pi_{12} = q^3 b_{10}^2 b_{01}$
$\pi_{13} = q^3 b_{10} b_{11} b_{01}$	$\pi_{14} = q(1-q)^2 b_{02}$	$\pi_{15} = q^2(1-q) b_{01}^2$	$\pi_{16} = 2q^2(1-q) b_{01}^2$
$\pi_{17} = 2q^2(1-q) b_{00} b_{02}$	$\pi_{18} = 2q^2(1-q) b_{02} b_{01}$	$\pi_{19} = 2q^3 b_{01}^2 b_{00}$	$\pi_{20} = q^3 b_{00} b_{01}^2$
$\pi_{21} = 2q^3 b_{01}^3$	$\pi_{22} = q^3 b_{02} b_{00}^2$	$\pi_{23} = 2q^3 b_{02} b_{00} b_{01}$	$\pi_{24} = q^3 b_{02} b_{01}^2$

TABLE V: Probabilities of observing different open triads originated from open triads

$\phi_1 = q^3 b_{20} a_{10}^2$	$\phi_2 = q^2(1-q) a_{10}^2$	$\phi_3 = q(1-q)^2 b_{20}$	$\phi_4 = 2q^2(1-q) b_{20} a_{10}$
$\phi_5 = q^2(1-q) a_{10} a_{01}$	$\phi_6 = q(1-q)^2 b_{11}$	$\phi_7 = q^2(1-q) b_{11} a_{10}$	$\phi_8 = q^2(1-q) b_{11} a_{00}$
$\phi_9 = q^2(1-q) b_{10} a_{01}$	$\phi_{10} = q^2(1-q) b_{11} a_{01}$	$\phi_{11} = q^3 b_{11} a_{10} a_{00}$	$\phi_{12} = q^3 b_{10} a_{10} a_{01}$
$\phi_{13} = q^3 b_{11} a_{10} a_{01}$	$\phi_{14} = q(1-q)^2 b_{02}$	$\phi_{15} = q^2(1-q) a_{01}^2$	$\phi_{16} = 2q^2(1-q) b_{01} a_{01}$
$\phi_{17} = 2q^2(1-q) b_{02} a_{00}$	$\phi_{18} = 2q^2(1-q) b02 a_{01}$	$\phi_{19} = 2q^3 b_{01} a_{01} a_{00}$	$\phi_{20} = q^3 b_{00} a_{01}^2$
$\phi_{21} = 2q^3 b_{01} a_{01}^2$	$\phi_{22} = q^3 b_{02} a_{00}^2$	$\phi_{23} = 2q^3 b_{02} a_{00} a_{01}$	$\phi_{24} = q^3 b_{02} a_{01}^2$

[5] J. B. Casterline, "Diffusion processes and fertility transition: Introduction," *Diffusion Processes and Fertility Transition*, pp. 1–38, 2001.

[6] H. Kohler, J. R. Behrman, and S. C. Watkins, "Social networks and HIV/AIDS risk perceptions," *Demography*, vol. 44, no. 1, pp. 1–33, 2007.

[7] C. A. Latkin, V. Forman, A. Knowlton, and S. Sherman, "Norms, social networks, and HIV-related risk behaviors among urban disadvantaged drug users," *Social Science & Medicine*, vol. 56, no. 3, pp. 465–476, 2003.

[8] E. D. Kolaczyk, *Statistical analysis of network data (Springer Series in Statistics)*. Springer-Verlag, 2009, vol. 69.

[9] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore, "On the bias of traceroute sampling: or, power-law degree distributions in regular graphs," in *Proc. 37th ACM Symposium on Theory of Computing*, 2005, pp. 694–703.

[10] A. Clauset and C. Moore, "Accuracy and scaling phenomena in internet mapping," *Physical Review Letters*, vol. 94, no. 1, p. 018701, 2005.

[11] F. Viger, A. Barrat, L. DallAsta, C. Zhang, and E. D. Kolaczyk, "What is the real size of a sampled network? the case of the internet," *Physical Review E*, vol. 75, no. 5, p. 056111, 2007.

[12] M. J. Salganik and D. D. Heckathorn, "Sampling and estimation in hidden populations using respondent-driven sampling," *Sociological Methodology*, vol. 34, no. 1, pp. 193–240, 2004.

[13] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in facebook: A case study of unbiased sampling of osns," in *Proc. INFOCOM*. IEEE, 2010, pp. 1–9.

[14] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proc. 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2006, pp. 631–636.

[15] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou, "Walking on a graph with a magnifying glass: stratified sampling via weighted random walks," in *Proc. ACM SIGMETRICS Intl. Conf. on Measurement and Modeling of Computer Systems*, 2011, pp. 281–292.

[16] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proc. 10th ACM SIGCOMM Conference on Internet Measurement*, 2010, pp. 390–403.

[17] P. V. Marsden, "Survey methods for network data," *The SAGE Handbook of Social Network Analysis*, pp. 370–388, 2011.

[18] ——, "Network data and measurement," *Annual Review of Sociology*, pp. 435–463, 1990.

[19] J. M. Perkins, S. Subramanian, and N. A. Christakis, "Social networks and health: A systematic review of sociocentric network studies in low-and middle-income countries," *Social Science & Medicine*, vol. 125, pp. 60–78, 2015.

[20] B. Wellman, "The community question: The intimate networks of east yorkers," *American Journal of Sociology*, pp. 1201–1231, 1979.

[21] C. S. Fischer, *To dwell among friends: Personal networks in town and city*. University of Chicago Press, 1982.

[22] J. R. Behrman, H. Kohler, and S. C. Watkins, "Social networks and changes in contraceptive use over time: Evidence from a longitudinal study in rural kenya," *Demography*, vol. 39, no. 4, pp. 713–738, 2002.

[23] S. Helleringer and H. Kohler, "Sexual network structure and the spread of hiv in africa: evidence from likoma island, malawi," *AIDS*, vol. 21, no. 17, pp. 2323–2332, 2007.

[24] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *New England Journal of Medicine*, vol. 357, no. 4, pp. 370–379, 2007.

[25] ——, "Social contagion theory: examining dynamic social networks and human behavior," *Statistics in Medicine*, vol. 32, no. 4, pp. 556–577, 2013.

[26] G. Harling and J. Onnela, "Impact of degree truncation on the spread of a contagious process on networks," *arXiv:1602.03434*, 2016.

[27] B. Fotouhi and N. Momeni, "Sampling and inference for fixed-choice ego-centric network design," 2016, working paper, DOI: 10.13140/RG.2.1.3453.1605/3.

[28] P. Hoff, B. Fosdick, A. Volfovsky, and K. Stovel, "Likelihoods for fixed rank nomination networks," *Network Science*, vol. 1, no. 03, pp. 253–277, 2013.

[29] A. Banerjee, A. Chandrasekhar, E. Duflo, and M. Jackson, "The diffusion of microfinance," *Science*, vol. 341, no. 6144, p. 1236498, 2013.

[30] K. M. Harris, C. T. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J. R. Udry, "The national longitudinal study of adolescent health: Research design," *Available at h ttp://www. cpc. unc. edu/pr ojects/addhealth/design*, 2009.

[31] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, pp. 1360–1380, 1973.

[32] N. E. Friedkin, "Information flow through strong and weak ties in intraorganizational social networks," *Social Networks*, vol. 3, no. 4, pp. 273–285, 1982.

[33] D. Krackhardt, "The strength of strong ties: The importance of philos in organizations," *Networks and Organizations: Structure, Form, and Action*, vol. 216, p. 239, 1992.

[34] M. Karsai, N. Perra, and A. Vespignani, "Time varying networks and the weakness of strong ties," *Nature Scientific Reports*, vol. 4, 2014.

[35] M. Burke and R. Kraut, "Using facebook after losing a job: Differential benefits of strong and weak ties," in *Proc. ACM Conf. on Computer Supported Cooperative Work*, 2013, pp. 1419–1430.

[36] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.

[37] S. G. Roberts, R. I. M. Dunbar, T. V. Pollet, and T. Kuppens, "Exploring variation in active network size: Constraints and ego characteristics," *Social Networks*, vol. 31, no. 2, pp. 138–146, 2009.

[38] M. E. J. Newman and D. J. Watts, "Scaling and percolation in the small-world network model," *Physical Review E*, vol. 60, no. 6, p. 7332, 1999.

[39] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[40] S. N. Dorogovtsev, P. L. Krapivsky, and J. F. F. Mendes, "Transition from small to large world in growing networks," *Europhysics Letters*, vol. 81, no. 3, p. 30004, 2007.

[41] P. Holme and B. J. Kim, "Growing scale-free networks with tunable clustering," *Physical Review E*, vol. 65, no. 2, p. 026107, 2002.

[42] G. Miritello, E. Moro, R. Lara, R. Martínez-López, J. Belchamber, S. G. Roberts, and R. I. Dunbar, "Time as a limited resource: Communication strategy in mobile phone networks," *Social Networks*, vol. 35, no. 1, pp. 89–95, 2013.

[43] R. I. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella, "The structure of online social networks mirrors those in the offline world," *Social Networks*, vol. 43, pp. 39–47, 2015.

[44] R. I. Dunbar, "The social brain: Psychological underpinnings and implications for the structure of organizations," *Current Directions in Psychological Science*, vol. 23, no. 2, pp. 109–114, 2014.

[45] A. Sutcliffe, R. Dunbar, J. Binder, and H. Arrow, "Relationships and the social brain: integrating psychological and evolutionary perspectives," *British journal of psychology*, vol. 103, no. 2, pp. 149–168, 2012.

[46] R. A. Hill, R. A. Bentley, and R. I. Dunbar, "Network scaling reveals consistent fractal pattern in hierarchical mammalian societies," *Biology Letters*, vol. 4, no. 6, pp. 748–751, 2008.

[47] W.-X. Zhou, D. Sornette, R. A. Hill, and R. I. Dunbar, "Discrete hierarchical organization of social group sizes," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 272, no. 1561, pp. 439–444, 2005.

[48] J. Saramäki, E. Leicht, E. López, S. G. Roberts, F. Reed-Tsochas, and R. I. Dunbar, "Persistence of social signatures in human communication," *Proceedings of the National Academy of Sciences*, vol. 111, no. 3, pp. 942–947, 2014.

[49] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994, vol. 8.

[50] M. H. Quenouille, "Notes on bias in estimation," *Biometrika*, vol. 43, no. 3/4, pp. 353–360, 1956.

[51] J. W. Tukey, "Bias and confidence in not-quite large samples," in *Annals of Mathematical Statistics*, vol. 29, no. 2, 1958, pp. 614–614.

CHAPTER 4

Additional Simulation Results for Network Sampling

4.1 Chapter Outline

In this chapter we present additional simulation results for the sampling setup of the previous chapter. The results presented in this chapter were not included in the published paper, but we add them here for the sake of completeness and to further investigate the performance of the estimators.

4.2 Introduction

Since the inference framework presented in the previous chapter was the first one to accommodate the properties of fixed choice surveys, we made certain simplifying approximations to enable us to calculate the results analytically. In this chapter, we focus on one of them whose effect we did not investigate in the paper. Namely, we focus our attention to the assumption of uniform response. We had assumed that when a seed node is asked to nominate B alters in the weak layer, the seed node chooses the B alters uniformly at random. In the following sections, we briefly investigate the effect

of three distinct mechanisms that can make the response process deviate from uniform selection.

4.3 Transitivity

In actual social networks, there are certain mechanisms that influence the connections. Transitivity is one of the most prominent factors: people tend to befriend the friends of their friends [Gra73]. This produces triangles and is the major cause of high clustering that is prevalent in actual social networks [WS98]. This transitivity tendency has already made its effect on the structure of the network, and our sampling scheme is taking place on the fixed structure in which the effect of transitivity is already incorporated. Inspired by this phenomenon, we consider one possible pathway of deviation from uniform response: transitivity of response.

Suppose individual X is friends with persons A and B, who are themselves friends with one another. Since human memory is associative [AB14], maybe if seed X remembers its friend A and mentions A, it is more likely to also subsequently remember B, as compared to some random neighbor. To incorporate this possibility, we consider the following basic scenario: the probability that seed X mentions node A is proportional to the number of mutual friends X has with A. Since this would assign zero chance for neighbors with whom X has no mutual friends, we use the idea of Laplace smoothing in machine learning [MRS08] and to construct the nomination probabilities, we add unity to the number of mutual friends.

Figure 4–1 shows the results for the performance of the estimators where the actual survey takes place with the above procedure, but the estimators employ the assumption of uniform response. It is visible from the figure that, as expected, the transitivity of response can cause bias in the results, but the accuracy is still within reasonable values.

4.4 Community Structure

Group identity is a strong and prevalent factor in social life, and group dynamics is a prominent topic that has always attracted much research in sociology and psychology [AM89, Taj81]. People universally exhibit ingroup favoratism. Various mechanisms, including peer pressure, social visibility and norm enforcement, norms of reciprocity and

trust, and conformity, drive individuals to bias their interactions towards the ingroup members as compared to outgroup members [Taj81, Gre14].

The group nature of social life can act as a pathway of deviation from uniform response. If seed X is connected to two neighbors A and B, where A belongs to the same social group as X (e.g., workplace, club, church, or neighborhood) and B does not, then X might be more likely to mention A as compared to B.

To model this community-driven response bias, we first need to generate a network which exhibits group structure. Group structure is often modeled in network science by employing modular networks constructed by stochastic blockmodels [KN11]. As a basic model, we used a stochastic blockmodel with two identical groups, where intergroup links are formed with a higher probability that intragroup links. Similar to Chapter 3, we imposed the constraint of realistic average degree.

We consider a basic setup in which ingroup neighbors have chance 0.7 and outgroup neighbors have chance 0.3. Figure 4–2 presents the results for the performance of the estimators. Again it can be seen that community-driven response introduces bias into the results, but in most cases the accuracy is within an acceptable range.

4.5 Popularity Bias

Social status is an important factor that drives interpersonal dynamics and interactions in humans and other social animals [CTSMM02]. Individuals carefully take into account their own and others' social status and plan their interactions accordingly [MSL87, MSLC01]. This is notable in every social group setting, both in adults and children [SS76]. There is also evidence that in surveys, there is a bias towards claiming to be friends with the popular individuals, creating a disproportionately many unreciprocated friendship claims towards these individuals [BN13].

We consider this phenomenon as a possible deviation pathways from uniform response. As a simple mode, we assume that each seed node chooses from its neighbors with probabilities proportional to their degrees. So a neighbor with a higher degree has a higher chance of getting mentioned by the seed node. Figure 4–3 presented the results for the performance of the proposed estimators. It can be seen that the resulting bias in this case is less than the two previous scenarios, and the estimators perform reasonably well.

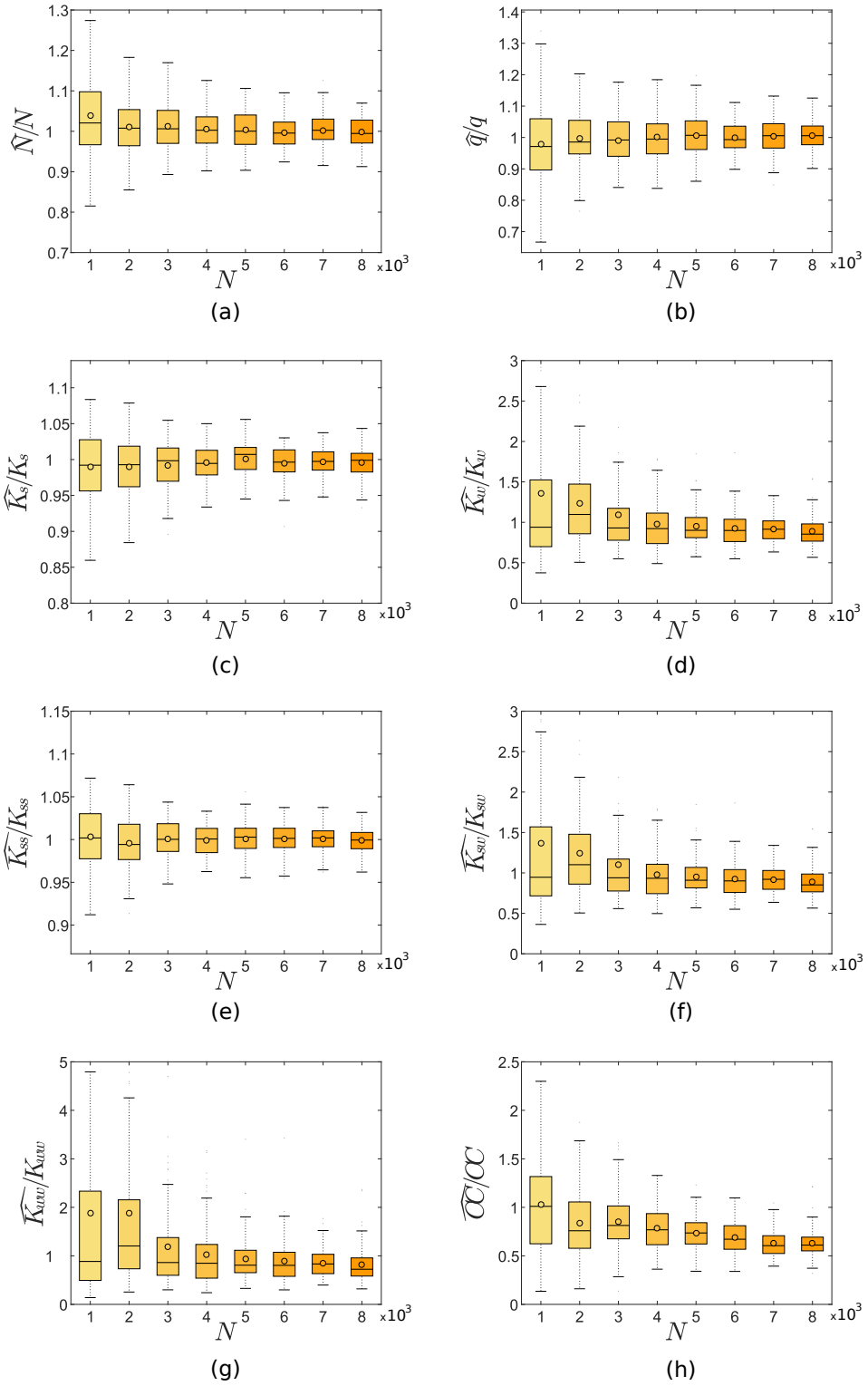


Figure 4–1: Results for transitivity-driven response procedure

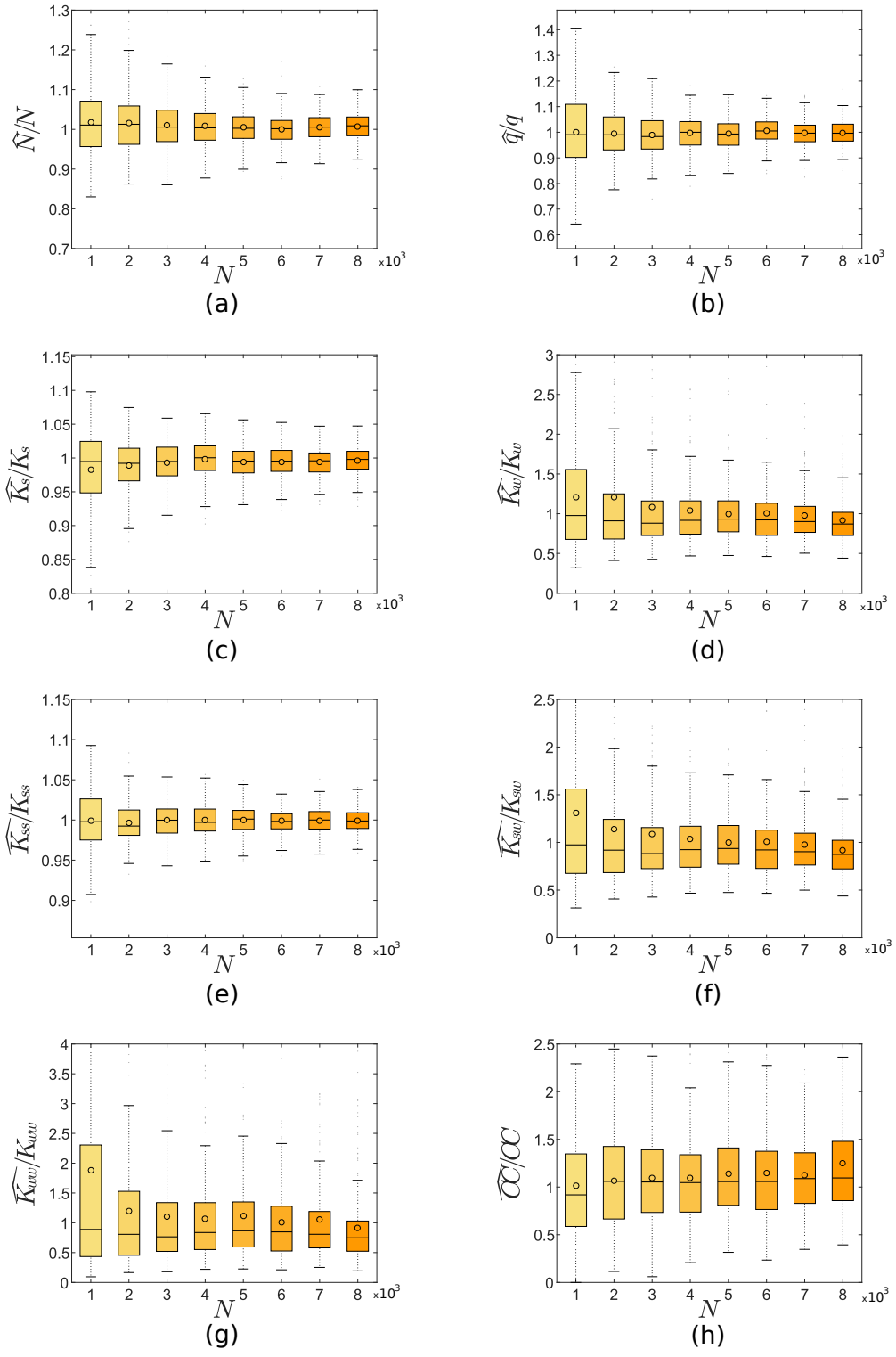


Figure 4–2: Results for community-driven response procedure

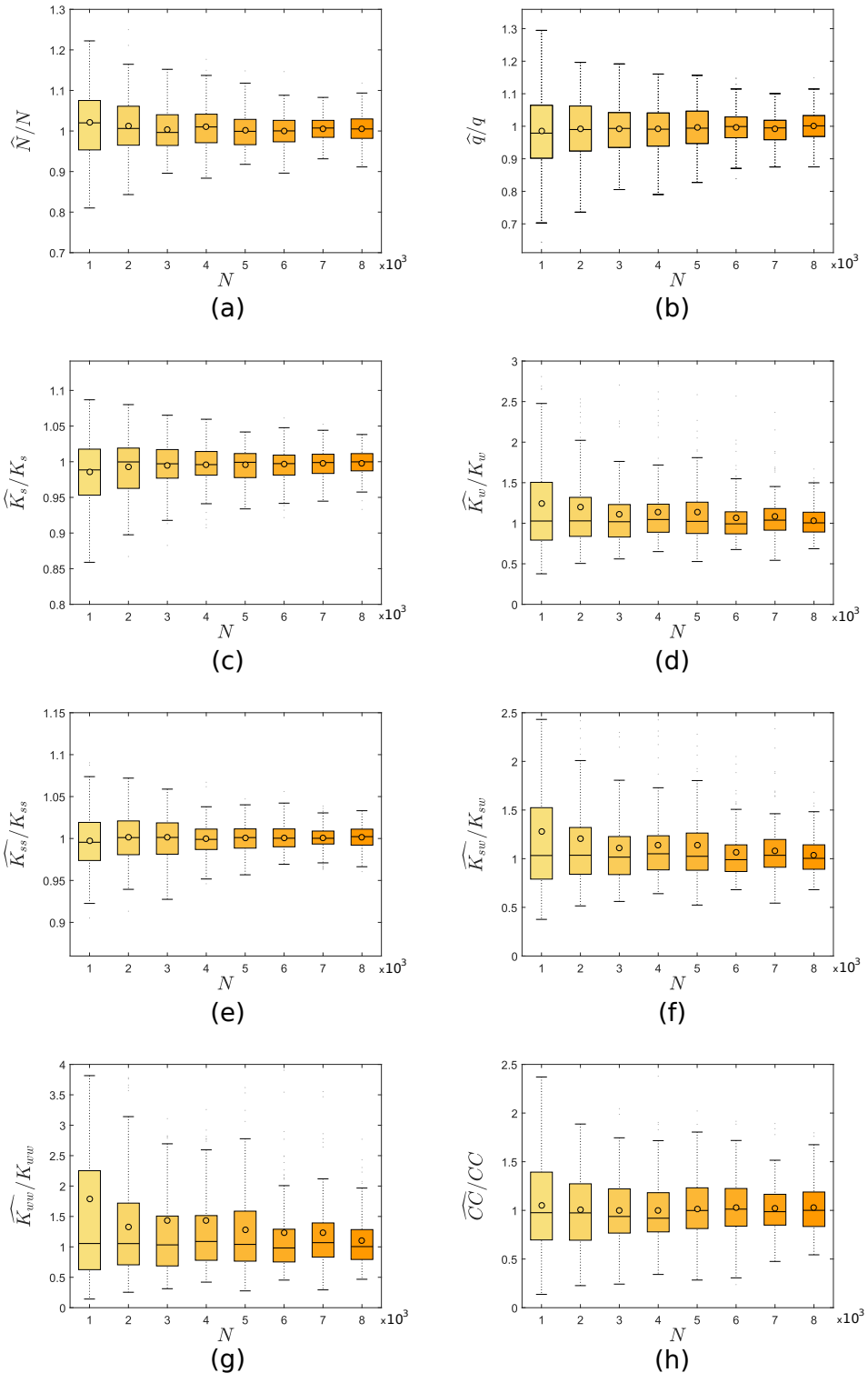


Figure 4–3: Results for popularity-driven response procedure

4.6 Chapter Summary

In this section, we saw that accommodating three possible deviation pathways from uniform response can introduce bias into the results, but in most cases the errors within reasonable limits. The resulting bias is the smallest for popularity-driven response, and is largest for transitivity-driven response. A possible extension to our paper can be incorporating these realistic response procedures in the sampling model and investigating the effects of their magnitude on the accuracy of the estimators.

CHAPTER 5

Introduction to Alter Sampling

5.1 Chapter Outline

In this chapter we introduce the concept of alter sampling. We first discuss some practical limitations in ascertaining the structure of offline social networks which would call for methods to utilize the structural features without having to observe the network. We then present examples of studies that have used alter sampling in practical settings with remarkable results.

5.2 Introduction

Not all network applications require knowledge of the global network structure. In some practical tasks, the highest priority might not be acquiring global knowledge to run subsequent analyses. Rather, we might need methods that can economically and quickly detect regions of the network with certain properties, or to target individuals with certain characteristics. For example, consider a village in which an epidemic outbreak is imminent, and vaccine resources are limited. Ideally, we would sample the whole network and use it in epidemic simulations to find the most efficacious set of nodes to be immunized. As we mentioned before, sampling the whole network is prohibitively costly. Even sampling will be too costly due to time limitations, because social network

studies often involve interviews and many survey questions. So we would need an efficient method to target well-connected individuals without having information about the global structure of the social network. As we discuss in Section 5.3.1, this is an example where alter sampling performs remarkable well.

Alter sampling is an efficient method for applications where acquiring global knowledge of the network structure is unfeasible, and we need to work with only local structural information. Below we discuss the theoretical study that first introduced this method in the context of network epidemiology. We then present a few recent studies that have succeeded in using this method in practical applications. In Chapter 6, we will demonstrate the effectiveness of this method across a wide range of social networks with diverse structural properties. Then in Chapter 7, we connect alter sampling to a social phenomenon that relates the popularity of individuals to those of their friends, and in Chapter 8 we use these information to shed light on the underlying organization of social networks and their structural inequalities.

5.3 Applications of Alter Sampling

In this section we discuss the practical applications of alter sampling. The study that introduced the technique is a theoretical one, and comes from the field of network epidemiology. Then we present four practical studies that have utilized alter sampling. The first one applies it to the friendship network in a dorm for early detection of flu outbreaks. The second study applies it to the Twitter for early detection of viral content. The third study applied it to Twitter for early forecasting Hurricane Sandy. The forth study utilizes alter sampling for health intervention in a public health program in Honduras.

5.3.1 Acquaintance Vaccination

As we mentioned above, there are practical settings where network sampling and inference might be infeasible. There might be situations where, due to time or budget constraints, or other practical concerns such as privacy issues or geographic dispersion, we cannot fully perform a sampling scheme that would give us multiple properties of the structure of the network. In this section, we illustrate this point with an example.

Consider the problem of vaccinating the individuals against some disease in a village, where the vaccine resources are limited and we have to choose a small fraction of the population for immunization. Naturally, we would require to choose the targets in a way that vaccinating them would give the highest herd immunity to the whole population. If we had complete knowledge of the structure of the underlying social network, we could run extensive simulations to help us devise optimal targets. Incorporating the specific features of the disease (transmissibility, recovery rate, immunity development, etc.), we could simulate the spread of the disease for all possible sets of vaccination targets and all possible initial sets of infected individuals, and choose the target set which has the minimum average outbreak size. Since complete knowledge of network structure is impractical, we could devise efficient sampling schemes to infer the structure. We would need to survey a fraction of the population, ask them to list their ties. Depending on the sampling scheme, we might also ask them to give further quantitative data for each tie, or we might ask them multiple follow-up questions about each alter they have named. There are many practical challenges to overcome in sampling social networks, and they are usually slow procedures by nature, because of the time that interviewers would require to spend with each respondent. Considerable time and resources would be needed for reliable sampling. Hence, we need to devise a strategy to choose the targets that is efficient, without requiring knowledge of the social structure. The fewer questions the strategy prompts us to ask the respondents, the better.

The naive vaccination strategy would be to randomly choose people and vaccinate the randomly-selected individuals. This would require little time, and would require asking no questions from selected individuals. However, that turns out not to be very effective in providing the most herd immunity possible. Intuitively, we need to find an efficient way to find the most well-connected individuals in the society. They interact with many people and if they get infected, they will transmit the disease to many people. Vaccinating them will be highly influential towards containing possible epidemic outbreaks. Random sampling does not necessarily capture these individuals, since it does not systematically target them. We cannot simply choose individuals with highest degree, nor can we choose them based on any other trait (e.g., how central they are in the network, etc.), because we do not know the structure of the network, thus we do not know structural traits of the individuals. We need a strategy that targets influential nodes without requiring knowledge of network structure.

In a theoretical study [CHBA03], using multiple classes of synthetic networks that emulate properties of real social networks, it is shown that a very effective strategy is randomly selecting individuals, asking them to name someone they know, and then vaccinating those that are mentioned. This scheme is called *acquaintance immunization* [CHBA03, MKC+04, GLA+07], and simulations show that despite their remarkably simple procedure, they are highly effective. The promising feature of acquaintance immunization is that it targets influential nodes with almost no knowledge of the structure of the underlying social structure.

This was the first time the idea of alter sampling was introduced. It is simple and yet highly effective. Below we move to studies that have successfully utilized this technique in practical settings.

5.3.2 Health Monitoring

Christakis and Fowler [CF10] conducted an empirical study using the idea of alter sampling. The study focuses on the problem of efficiently monitoring the spread of the H1N1 flu. The authors hypothesize that neighbors of randomly-selected nodes get infected on average earlier than the whole population. To test this hypothesis, the authors monitored the spread of flu in Harvard College for 4 months. They compared two distinct groups of students, one acquired through random sampling from students, and the other consisting of those who were named by others as friends. In the terminology we used for acquaintance optimization, the first group are the result of random sampling and the second are the result of alter sampling. The authors observe that the prevalence curve (number of flu cases as a function of time) for the latter group is shifted 13.9 days forwards in time as compared to the former group. The authors highlight that this can be utilized for the detection of outbreaks in the early stages of an epidemic. Interestingly, the authors show that if instead of being nominated as a friend by some other individual, we use the subjects' self-reported claims of popularity, no significant shift is detectable. This suggests that claiming to be well-connected does not necessarily indicate that this is actually true. We will return to this disparity later in Chapter 7.

5.3.3 Information Diffusion on Social Media

The idea of using alter sampling for the early detection of outbreaks can be extended beyond infectious diseases, and can also be applied to information contagion. In [GHMC+14], the global diffusion of viral online content on Twitter is studied. In Twitter, each piece of information that a user posts is called a tweet (if the user generates it) or retweet (if the user shares a post someone else has generated), and topics can be marked by hashtags (so that one can search for a hashtag and get the pertinent tweets).

The typical size of data extracted from online social networks is orders of magnitude greater than offline social networks. In [GHMC+14], for example, the social network that is used contains 1.5 billion ties and half a billion messages during the period of the study. The hypothesis is that small samples of users obtained by alter sampling (these users are refered to as *sensors*) receive the viral hashtags quicker than samples obtained by random sampling (similar to the flu case, where the outbreak peaked earlier for the former group). In [GHMC+14], the authors compared many samples obtained from either sampling method, and confirm that alter sampling gives samples that lead in the access to viral hashtags. One possible issue remains, as the authors of [GHMC+14] point out and then control for: perhaps it is not network position that determines access to viral content, but the converse is true. That is, perhaps it is not the case that sensors get access to viral content because they are somehow central, but instead, perhaps they are central *because* they always produce interesting or important content that goes viral frequently, so they attract many followers and become central as a consequence. This is ruled out by accounting for the generation of the content, and the authors show that the amount of viral content generation of these nodes is not significantly higher than average. Moreover, the authors of [GHMC+14] point out that there is a positive association between centrality and number of tweets. The authors rule out a second confounding possibility: perhaps merely because these people tweet more often, the viral tweets happen to show up in their tweets more often than other users. The authors rule out this possibility by a shuffling test: they fix the number of tweets, and randomly redistribute all the tweets among users. If we still observe a lead in the access time, this is the sole effect of excess tweets. The authors observe a small lead after the shuffling test, and argue that the difference between this observed lead

and the one originally observed is the effect that pertains to genuinely-quicker access of these users to viral content. These users, because of their position in the network, have an earlier exposure on average to viral content. The authors contend that this can be used for early detection of viral online content. The authors remark that the early detection of global mood and taste patterns have potential applications in marketing and policy making.

5.3.4 Early Detection of Natural Disasters

An information-diffusion approach was taken in [KCM+15] for the early detection of natural disasters. In [KCM+15], the authors have a random sample of Twitter users as the control group, and one acquired via alter sampling as the sensor group. The authors are interested in the lead-time of awareness in the sensors as compared to the control group. The authors also use the geo-locations of tweets and geographic data from the National Hurricane Center. They show that the hurricane-related content (blackout, weather change, etc.) appears in the tweets of the sensor group on average 11 hours earlier than the control group. Furthermore, the authors show that if the sensor data is combined with geo-location data, the lead time increases to 26 hours. This is clearly a considerable opportunity for practical purposes. Such temporal lead can help individuals for quicker preparation against an incoming hurricane and to swiftly learn about possible consequences (such as blackouts or property damages) from those who have experienced it earlier and take necessary safety measures.

5.3.5 Health Intervention

The health-intervention applications of alter sampling are not limited to vaccination or early detection of epidemics. It can be used to promote the spread of information and awareness in social networks. In a recent study published in The Lancet, this method is used for improving the impact of health intervention in 32 villages in Honduras [KHS+15]. Two distinct health interventions (one nutritional and the other pertaining to water purification) were made. In villages, the products and instructions were given to 5% of the population which were randomly selected. In some other villages, the targets were chosen via alter sampling. The final prevalence of the adoption

of the health behaviors were greater where alter sampling was used (about 12%). Also, the general level of knowledge about the health behaviors was higher at the end of the study period in those villages.

5.4 Chapter Summary

In this chapter we introduced the idea of alter sampling and the situations in which alter sampling would be considerably helpful. We discussed its origins in the network epidemiology literature and its subsequent adoptions in practical studies for early detection of epidemic outbreaks, information diffusion in online social media, natural disasters, and promoting public health intervention policies.

CHAPTER 6

Paper: Effectiveness of Alter Sampling in Various Social Networks

The material presented in this chapter is submitted to Scientific Reports.

Please note that the references of the manuscript are listed at the end of this chapter.

Robustness of Alter Sampling in Social Networks

Naghmeh Momeni and Michael G. Rabbat

Abstract

Social networks have a key role in studying various individual and social behaviors. To use social networks in a study, their structural properties must be measured. For offline social networks, the conventional procedure is surveying/interviewing a set of randomly-selected respondents. In many practical applications, inferring the network structure via sampling is too prohibitively costly. There are also applications in which it simply fails. For example, for optimal vaccination or employing influential spreaders for public health interventions, we need to efficiently and quickly target well-connected individuals, which random sampling does not do. In a few studies, an alternative sampling scheme (which we dub 'alter sampling') has proven useful. This method simply targets randomly-chosen neighbors of the randomly-selected respondents. A natural question that arises is how generalizable this method is. Is the method suitable for every social network or only the very few ones considered so far? In this paper, we demonstrate the robustness of this method across a wide range of networks with diverse structural properties. The method outperforms random sampling by a large margin for a vast majority of nodes in all the networks. We then propose an estimator to assess the gain of choosing alter sampling over random sampling in practical scenarios, and demonstrate its accuracy via Monte Carlo simulations on diverse synthetic networks.

INTRODUCTION

Social networks are mathematical tools for modeling social relations and interactions, and for studying the interplay between structure and agency. They are employed in studying various social phenomena, such as contagion of health behaviors and the adoption of new ideas and behaviors [1]–[3], the spread of infectious disease [4], [5], the diffusion of

information [6], [7], and the effect of network position and connections on individuals' power [8], [9], job opportunities [10], cooperation [11], mental health [12], longevity [13], behavioral and ideological influence [14]–[16], and migration decisions [17], [18].

Descriptive studies of social networks relate the observed behavior of a social dynamical process or individual trait to the structural properties of the social networks. Recent studies also seek to leverage the theory of social networks for practical applications, such as 'seeding' strategies and finding influential spreaders [15], public-health interventions [19], and for early detection of epidemic outbreaks [20]. This paper focuses on a specific practical method, which we call 'alter sampling', that economically targets influential nodes while remaining agnostic of the network structure. We first briefly review a few successful applications. We then provide a case study on using alter sampling on various social networks with different structural properties, and show that it performs remarkably well in all of them. Finally, we propose estimators for the gain in using alter sampling over random sampling. We conclude by discussing the implications of the effectiveness of alter sampling on how social networks are organized.

In social network studies, descriptive or practical, analysis is carried out in terms of standard 'network statistics', i.e., quantities that pertain to the structural properties of the social networks (e.g., degree, measures of centrality, clustering, homophily). These properties need to be observed and measured first. Unlike some networks with non-social origins (e.g., the Internet and the World Wide Web), measurements in offline social networks are costly and challenging. Efficient sampling and inference methods are needed to meet the specific challenges of social networks.

In practice, there are situations where, due to time or budget constraints, or other practical concerns, a sampling procedure would be unfeasible. As an illustrative example, consider the problem of vaccinating individuals against some disease in a village, where the vaccine resources are limited and we have to choose a small fraction of the population for immunization. It would be ideal to have complete knowledge over the network structure to identify the targets optimally. Considerable time and resources would be needed for acquiring such complete knowledge of the network structure. It is practically implausible.

So we need to devise an efficient strategy to identify the targets without requiring knowledge of the social structure. The fewer questions the strategy required us to ask the respondents, the better.

The cost-effective, but naive vaccination strategy would be to randomly choose individuals for vaccination. Intuitively, we need to find an efficient way to find and vaccinate the well-connected individuals, because if they get infected, they will transmit the disease to many people. Random sampling does not necessarily capture these individuals because it does not systematically target them. It is shown that a very effective strategy is randomly selecting individuals, asking them to name someone they know, and then vaccinating those that are mentioned. This scheme is called *acquaintance immunization* [22]–[24], and simulations show that despite its remarkably simple procedure, it is highly effective. In this paper, to use a more broader term that is also applicable to non-epidemics contexts, we use the term *alter sampling* to refer to the method of random selection of neighbors of a random sample.

The promising feature of alter sampling is that it targets influential nodes with almost no knowledge of the structure of the underlying social structure. Christakis and Fowler [20] describe an empirical study using the idea of alter sampling to monitor the spread of the H1N1 flu. Comparing two samples of students, one obtained via random sampling and one via alter sampling, they showed that the prevalence curve for the latter sample is shifted 13.9 days forwards in time as compared to the former. This indicates that alter sampling can be utilized for the detection of outbreaks in the early stages of an epidemic.

The idea of using alter sampling for the early detection of outbreaks can be extended beyond infectious diseases, and can also be applied to information contagion. The diffusion of viral online content on Twitter is an example, where it is shown that samples of users obtained by alter sampling (refered to as *sensors*) receive viral hashtags earlier than samples obtained by random sampling. The difference still remains after controlling for possible reverse causality (that sharing viral content is is not the result, but the cause of network position) by showing that virality of posts and network position are not significantly correlated [21].

Alter sampling can also be used to promote the spread of information and awareness in social networks. In a very recent study, this method was used for improving the impact of health intervention in 32 villages in Honduras [19]. Two distinct health interventions (one nutritional and the other pertaining to water purification) were made. The products and instructions were given to 5% of the population (reached via random sampling in some villages and via alter sampling in others). The final prevalence of the adoption and the general knowledge of the health behaviors were greater in villages where alter sampling was used (about 12%).

The above empirical observations suggest that alter sampling is a potent and efficient practical method for finding influential nodes. To be able to confidently use it in practice, we need to verify that the success of the method was not due to peculiarities of the above (very few) cases, and to ascertain its robustness across a wide range of networks with social origin. This is the first focal task of the present paper. We demonstrate that alter sampling is robust in a range of networks with social origins with diverse structural properties (we consider positively, negatively, and neutrally assortative networks, high and low degree variance, different levels of clustering and density). We demonstrate that alter sampling performs well across all of them, and performs remarkably similarly. This sheds light on micro mechanisms that are present in networks with social origin that do not depend on the specific properties of the context.

A major practical issue to consider when employing alter sampling is how advantageous it is. That is, we need to quantify the benefit of using alter sampling as compared to random sampling. Since we are considering scenarios in which the structure of the social network is unknown, we cannot use any structural information to estimate the benefit of using alter sampling, either a-priori or retrospectively. For example, after using interview data to vaccinate the alters that respondents nominate, how can we assess the gain of this method over the random method? Answering this question is the second focal task of the present paper. We propose estimators that use interview data to quantify the ‘gain’ of alter sampling.

RESULTS

We can quantify the performance of alter sampling in various ways. The most basic individual attribute that characterizes the influence of a node on dynamical processes on networks is the degree. Thus we take the expected value of the degrees of the nodes reached via a sampling scheme as its merit, and the gain of choosing one sampling scheme over another is quantified as how this merit changes. With the few recent exceptions of employing alter sampling in practical settings, most studies have employed random sampling. In the present paper, we seek to quantify how much loss is associated with such a decision. For undirected networks, we compare the degree of node x with the mean degree of its neighbors. The ratio of these two quantities is the local gain of choosing alter sampling over random sampling. For node x, we denote this ratio by $\mathcal{G}_x$. The average value of this ratio over all nodes, which we denote by $\mathcal{G}$, gives the expected gain. Choosing alter sampling is justified if $\mathcal{G} > 1$. In directed networks, if node y follows node x (that is, there is a link from node y to node x), then y is called the in-neighbor of x, and x is called an out-neighbor of node y. The number of in-neighbors and out-neighbors of a node are called its in-degree and out-degree, respectively. For directed networks, we only consider in-degrees, because social influence operates in the direction of links. That is, for a given node x, it is the in-neighbors of x that are influenced by x, and not the out-neighbors. Otherwise, the methodology is similar to the case of undirected networks.

The descriptions of the data sets used are provided in the Methods section. their summary statistics are presented in Table I (for directed networks) and Table II (for undirected networks).

There are various ways we can quantify how useful alter sampling is, that is, for what proportion of the population it would be better to reach their neighbors via alter sampling as compared to reaching themselves via random sampling. In Figure 1 and Figure 2, we depict the proportion of nodes in each degree-percentile for whom $\mathcal{G}_x > 1$ in directed and undirected networks, respectively. It can be observed that for all networks, a vast majority of nodes do meet this criterion. For all directed networks under consideration, over 85% of the population have $\mathcal{G} > 1$. For undirected networks, this number is even higher (near

95%). Note that this phenomenon holds whether we compare the degree of each node with the mean or with the median degree of the neighbors, which highlights the robustness of the observed phenomenon against possible outliers.

So far we demonstrated that the observed seemingly-universal gain of alter sampling in social networks is not attributable merely to hubs. The above measures for prevalence of $\mathcal{G} > 1$ indicate that for a high proportion of the population, alter sampling is superior to random sampling, but these measures do not quantify to what extent that is so. To investigate this, we plotted the histogram of $\mathcal{G}$ across all nodes for different networks in Figure 3. It can be seen that across all these networks, the distribution of $\mathcal{G}$ is highly skewed, that is, there exist nodes for which the gain of using alter sampling is overwhelmingly large, and for the majority of nodes this gain is still considerably large (that is, $O(10^1)$ gain for alter sampling). As before, the gain is robust against outliers, as using median instead of the mean to define the gain does not alter the results significantly.

Furthermore, in Figure 4, we plotted the average $\mathcal{G}_x$ value of nodes as a function of their degree percentile. It can be seen that in all networks, the gain is considerably high for a vast majority of nodes. Note that the behavior of the gain function is similar across all networks, whereas their structural properties (such as assortative mixing, clustering, density, average degree, and degree variance) are widely different, as reported in Table I and Table II. This suggests that alter sampling is considerably robust against variation of network structure. This endows alter sampling with a notable versatility, thus it can be reliably used in practice for cases where it is not feasible to obtain the structure of the underlying social network through standard methods of social network studies, such as interviewing and surveying the population.

Now we attend to an important practical question, that is, to estimate the gain of choosing alter sampling over random sampling from empirical data. We provide two distinct estimators for the gain in choosing alter sampling over random sampling. In practice, an offline social network study (such as those that would be needed for optimal vaccination) involves interviewing people and asking them to nominate alters. Suppose that we also ask people to report the number of people they know [27]. We aim to estimate the gain of alter

	GitHub	Pokec	Twitter
N	46423	531478	5489933
E	156280	30622564	193245641
$\overline{k}$	3.366	18.754	35.2
$\hat{k}_{in}$	1	8	4
σ_{in}	20.25	32.140	989.01

TABLE I: Directed networks: N and E are number of nodes and number of edges, respectively. $\overline{k}$, $\hat{k}_{in}$ and σ_{in} denote average in-degree (which is equal to average out-degree), median of in-degrees and standard deviation of in-degrees, respectively.

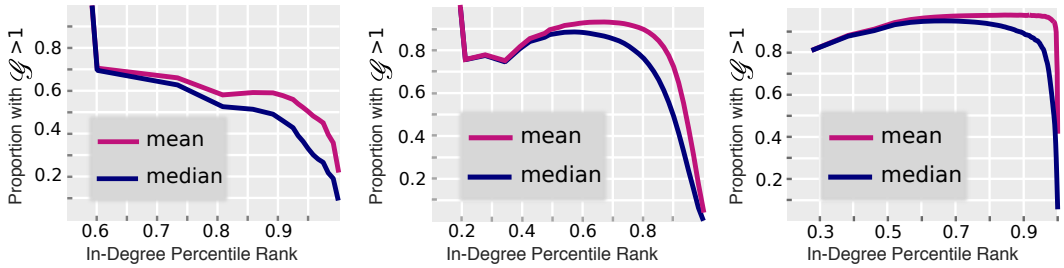
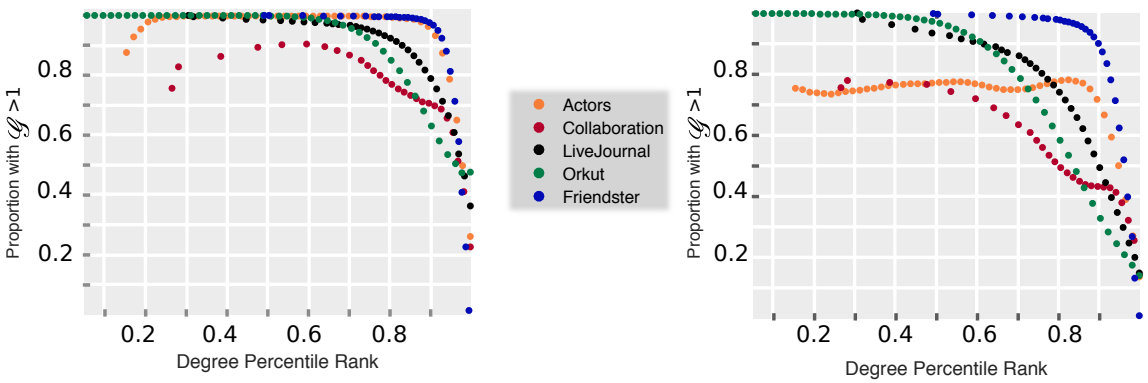


Fig. 1: Empirical distribution of nodes with $\mathcal{G}_x > 1$ as a function of in-degree percentile rank for directed networks.

sampling from the sequence of degrees that the respondents provide. Note that we cannot ask a respondent about the degree of the alter, because although people generally have a good knowledge of their own social ties, they might not be necessarily good at providing reliable estimates for the number of ties of one of their friends. Suppose that the underlying network has degree distribution $p(k)$. This means that a randomly-chosen respondent has degree $p(k)$. Suppose that a node with degree ℓ is mentioned as an alter. There are $Np(\ell)$ nodes of degree ℓ in the network, where N is the network size. Each of these nodes have ℓ neighbors that could be the initial respondent. Thus, there are on average $N\ell p(\ell)$ nodes that could have mentioned a degree-ℓ node. Denoting the mean degree of the network by μ_1, the probability that a mentioned alter has degree ℓ is given by $N\ell p(\ell)/\sum_\ell N\ell p(\ell) = \ell p(\ell)/\mu_1$.

	Actors	Collaboration	LiveJournal	Friendster	Orkut
N	894615	69032	3997962	22493449	3072441
E	57060378	450622	34681189	180606713	11785083
$\overline{k}$	127.5	13.05	17.40	16.058	76.28
$\widehat{k}$	41	5	6	3	45
σ_k	317.5	27.97	42.95	53.29	154.78
$r_{kk'}$	0.20	0.6018	0.045	-0.1816	0.0158
$\overline{C}$	0.4724	0.5977	0.2842	0.0734	0.1666

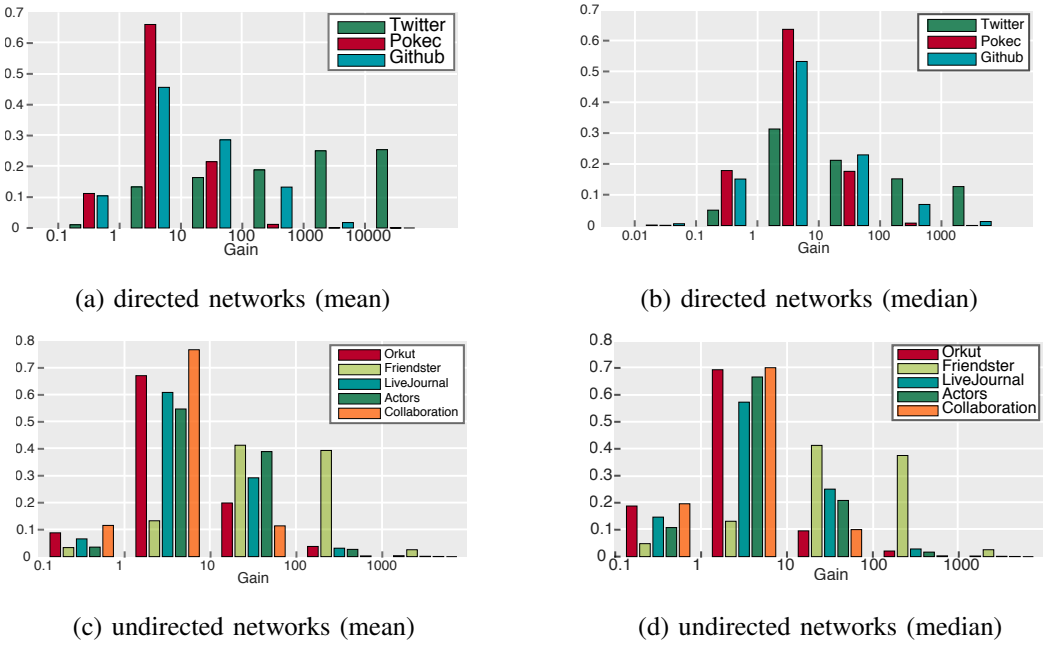
TABLE II: Undirected networks: N and E are number of nodes and number of edges, respectively. $\overline{k}$, $\hat{k}$ and σ_k denote average degree, median of degrees and standard deviation of degrees, respectively. Degree assortativity and average clustering coefficient of the graph are denoted by $r_{kk'}$ and $\overline{C}$, respectively.



(a) mean (b) median

Fig. 2: Empirical distribution of nodes with $\mathcal{G} > 1$ as a function of degree percentile rank for undirected networks. In (a), gain is defined using the mean, and in (b), using the median, as discussed in the text.

We use these conditional probabilities to obtain the expected degree of a named alter, which is $\sum_\ell \ell^2 p(\ell)/\mu_1 = \mu_2/\mu_1$, where μ_2 is the second moment of the degree distribution. For



(a) directed networks (mean)

(b) directed networks (median)

(c) undirected networks (mean)

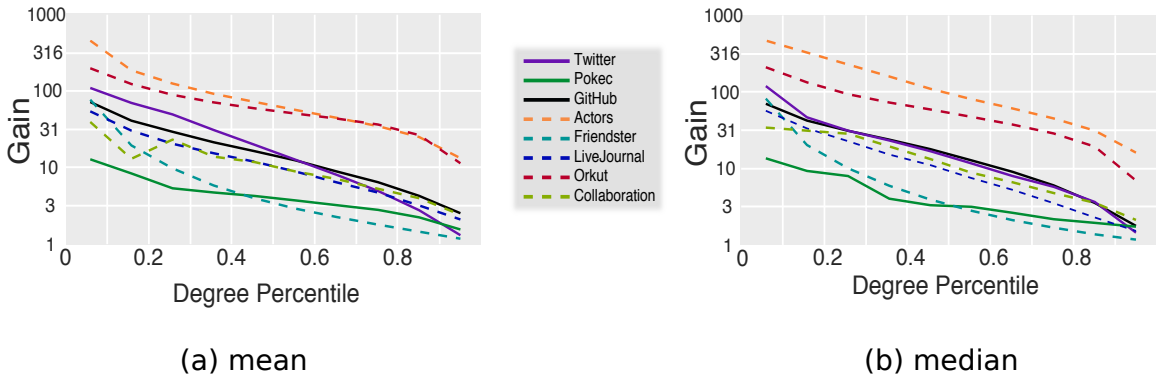
(d) undirected networks (median)

Fig. 3: The distribution of $\mathcal{G}$ for different networks. The top row pertains to the directed networks and the bottom row pertains to the undirected networks. In the left column, the gain for each node is defined as the average degree of its neighbors to its own degree. In the right column, the median degree of the neighbors is used instead of the mean.

node x, the gain of alter sampling is simply the ratio of the expected alter degree to that of node x. Thus the expected gain is given by $(1/N)\sum_x[(\mu_2/\mu_1)(1/k_x)]$. Let us denote the harmonic mean of the degrees by μ_h. That is, we have $\mu_h = \sum_k k^{-1}p(k)$. Thus, the expected gain is given by $\mathcal{G} = \mu_2\mu_h/\mu_1$. Denoting the set of respondents by $\mathcal{R}$, the total number of respondents by r, and the reported degree of respondent i by $\tilde{k}_i$, we arrive at the following estimator for the gain of choosing alter sampling:

$$\widehat{\mathcal{G}} = \frac{\left(\sum_{i\in\mathcal{R}} \tilde{k}_i^2\right)\left(\sum_{i\in\mathcal{R}} \frac{1}{\tilde{k}_i}\right)}{r\sum_{i\in\mathcal{R}} \tilde{k}_i}. \tag{1}$$

To assess the performance of the estimators, we would technically need an empirical sampled networked data set via the above mechanism for which the real underlying network is also known. Every existing offline social network data set in the literature is already the



(a) mean (b) median

Fig. 4: The gain of performing alter sampling as a function of the degree percentile of the target node. In (a), the average degree of neighbors is compared to the degree of each node to define the gain, and in (b), the median is used.

sampled version, and for none of them the true underlying web of interactions between people is known. Noting that this caveat prevents testing the estimators on real networks, we use synthetic networks. We choose network models that are proposed in order to emulate the properties of real social networks. One such model is the small-world network model [28]. This model was proposed in order to capture two important structural features observed widely in real social networks: high clustering, and small average path length. The former captures the high transitivity that is typical in networks of social origin (that is, friends of a person tend to become friends with high probability), and the latter pertains to the well-known six-degrees-of-separation phenomenon (that every two persons in society are connected via a very short chain of acquaintances). We synthesized 10000 networks (see Methods for details of the generation process of all network models considered) and for each case we estimated the gain from the above estimators and calculated its ratio to the true value. The closer this ratio is to unity, the better the estimators are performing. Figure 5a presents the results. It can be observed that the estimators are performing with acceptable accuracy, with errors mostly less than 10%.

The second conventional network generation model is the preferential attachment model. Proposed in [33] and later in [34], this model emulates the empirically-observed heavy-tailed nature of the degree distributions in diverse networks (such as the network of scientific citations, scientific collaborations, and the worldwide web). The results for this model are presented in Figure 5b.

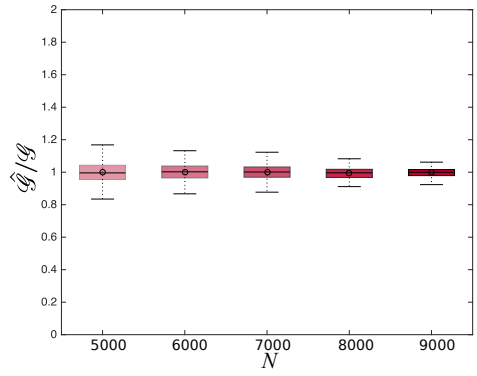
The third generative network model that we use is the one proposed in [35]. The model combines the preferential attachment model of network growth with high clustering. We refer to this model as the HK model. The HK model adds a triad-formation step to the conventional preferential attachment model, and makes it more suitable to modeling networks of social origin than the basic preferential attachment model (which has vanishing clustering coefficient for large networks). The results for the HK model are presented in Figure 5c. It can be observed that the variance of the estimator is slightly higher than it was for small-world networks, but it is slowly decreasing with network size.

The fourth model, proposed in [37] (which we refer to as the KE model), in addition to high clustering and skewed degree distribution, yields small average path length. Similar to the previous models, the estimator has an error of less than 10% in the majority of the simulation trials. The results for the KE model are presented in Figure 5d.

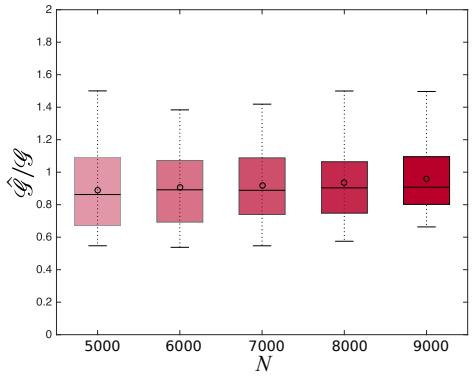
In all the simulation trials, the fraction of randomly-chosen respondents (whose random neighbors then constitute the alter set) are chosen uniformly at random between 0.1 and 0.2. So, equivalently, the value of r in Equation (1) is randomly selected between 10% and 20% of the total population.

DISCUSSION

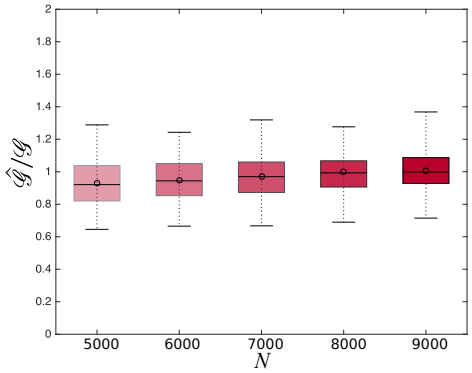
The presented results suggest that alter sampling is a strong and economical method for targeting well-connected nodes in the network when standard sampling procedures are costly and infeasible. The results demonstrate a remarkable versatility and robustness of this method. We considered many network data sets with large size and diverse structural properties (positive, negative, and neutral assortative mixing, high and low clustering and density, high and low variance of degrees and average degree), and in all cases, alter



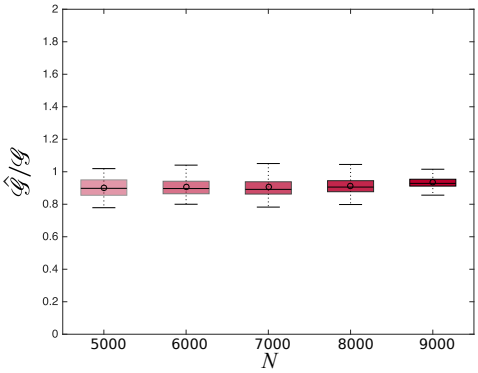
(a) small-world model (high clustering, low average path length)



(b) preferential attachment model (low clustering, heavy-tailed degree distribution)



(c) HK model (high clustering, heavy-tailed degree distribution)



(d) KE model (high clustering, low average path length, heavy-tailed degree distribution)

Fig. 5: The performance of the first proposed estimator for the gain of alter sampling for different families of networks.

sampling is advantageous over random sampling for a vast majority of nodes. This holds even if we consider the median of the degrees of neighbors instead of the mean in order to define the gain of choosing alter sampling over random sampling. Hence, although in the literature this phenomenon has been linked to the presence of hubs, our results indicate that a more prevalent structural feature must exist in all these networks to give rise to this behavior. Our results suggest that a continuous hierarchical structure must be present in all

these networks of social origin. In this hierarchy of degrees, every node either connects up or across, and nodes seldom connect down. This means that, nodes with very few links connect to both medium-connected and well connected nodes, as well as other weakly-connected nodes similar to themselves. Nodes with medium degrees connect to nodes with high degrees as well as other nodes with medium degrees. Nodes with high degrees only connect to other nodes with high degrees. This extends to the most highly-connected nodes in the entire network. This pattern exists in both directed and undirected networks considered. This suggests that networks with social origin might exhibit features that are macro outcomes of micro mechanisms which might be universal characteristics of human social behavior.

We also proposed an estimator to assess the gain of choosing alter sampling over random sampling in practical scenarios and investigated its performance on synthetic networks generated via four distinct conventional network generation models. We observed that the proposed estimator performs remarkably well across a diverse range of structural parameters of the synthetic networks.

The immediate extensions of the problem setup to more practical scenarios would be to consider imperfect response (due to, for example, forgetting or fatigue). Also, in some cases it might not be feasible to ask respondents to count the number of their friends. It is time-consuming and there might be situations in which there is only time to ask a few alter names. In this case, we will have to estimate μ_2, μ_1, and μ_h from the response data. Usually, there is a cutoff on the number of alters each respondent must mention, which is typically less than 10. In this case, the above moments of the degree distribution must be estimated from a dataset in which for each node only about 10% of the links are known. This is an interesting problem of statistical inference with immediate practical importance. We hope the results presented in this paper will invite closer investigations of alter sampling and its robustness and limitations, as well as the associated network sampling problems that will be practically imperative.

METHODS

Network Models

Small-world: We use a variant [29] in which the network is built as follows: we begin with a $2b$-regular lattice (a ring in which each node is connected to b immediate neighbors from each side), and we create each non-existing link with constant probability p, independently. Since it has been consistently shown in the literature that cognitive constraints limit the effective number of social ties a human can actively maintain to about 150 [27], [30]–[32] (also called the *Dunbar number*), we restrict the space of parameters to a domain for which the average degree is about 150. The value of b was randomly chosen between 5 and 10, and the value of p was chosen in a way to yield the average degree no greater than 200.

Preferential Attachment: In this model, nodes are added to the network sequentially, and each incoming node attaches to m existing nodes that are selected with degree-proportional probabilities. We selected m randomly between 50 and 75, generating networks with average degree between 100 and 150. We considered sizes from 5000 to 9000. We synthesized 1000 networks for each size.

HK: The parameters of the model are the initial number of links that each incoming node creates when it is being added to the network, and the triad formation probability. In the ensemble of networks that we generated, we randomized the first parameter between 50 and 100, and the triad formation probability was randomly generated in the interval $[0, 0.5]$, and a network was only accepted if the mean degree was less than 150. For each network size, we generated 1000 synthetic networks and implemented the sampling procedure described above with G randomly chosen between 5 and 10, because values of G more than 10 are rare in real social network studies [36].

KE: In this model, at each timestep these are m active nodes and as a new node is added, it creates m links. Each link, with probability μ connects to a random node chosen with degree-proportional probabilities according to the basic preferential attachment scheme, and with probability $1 - \mu$ attaches to one of the active nodes. The new node becomes

active and one of the previously-active nodes becomes inactive with probabilities inversely proportional to degrees. This procedure is then repeated. We have randomized the parameter space with the restriction that the generated networks have mean degree between 100 and 200.

Data

To ascertain the versatility of alter sampling, we considered five undirected and three directed networked data sets. A quantitative summary of their properties is presented in Table (I) and Table II, for directed and undirected networks, respectively. Below we provide a qualitative description of the data sets:

Film Actor Network: We use a network derived from the IMDB movie/actor network available in the University of Florida Sparse Matrix Collection [43]. This bipartite network consists of 428,440 movies and 896,308 actors and stores the movies in which each actor has appeared. Based on this graph, we can build the co-starring network. In this network each node represents an actor and an edge connecting two nodes indicates that those nodes have co-appeared in at least one movie. Note that we do not consider weights for the edges.

Scientific Collaboration Networks We use the collaboration network available at [38]. The dataset is extracted from the e-print arXiv and covers scientific collaborations between authors papers in five categories in the period from January 1993 to April 2003. If an author x co-authored a paper with author y, the graph contains a undirected edge from x to y.

LiveJournal LiveJournal is a social networking service where users can keep a blog, journal or diary and also can declare friendship with each other. The network that we use here is available at [39], [40] and consists of about four million users.

Friendster Friendster is an on-line gaming network. Before re-launching as a game website, it was a social networking site. The network that we use in this paper is a subset of the graph available at [41] consisting of more than 22 million nodes.

Orkut Orkut is a free on-line social network. The network used in this paper is available

at [41] and consists of more than three million users.

Twitter For the network of Twitter users, we use the dataset collected by Kwak et al. [42]. This dataset describes the connectivity among users who joined Twitter prior to August 2009. The subgraph that we use has 5.8 million users and more than 193 million edges.

GitHub The site github.com offers free code repository hosting for public projects and paid code repository hosting for private projects. Individuals can follow one another, like users of Twitter, in order to stay aware of each other's activities. In [44] the GitHub Archive site[1] was used to download past compressed archives of hourly activities over a one-year period. The collected and processed data are used to create multiple graphs including the followership graph which is used in this paper. In this graph there is an edge from node x to node y, if user x follows user y.

Pokec Pokec is the most popular on-line social network in Slovakia. The dataset is available at [45] and consists of more than 1.6 million nodes and more than 30 million edges.

AUTHOR CONTRIBUTIONS STATEMENT

N.M. and M.R. conceived the research problem. N.M. gathered the data and performed the analyses. N.M. and M.R. discussed the results, and wrote and reviewed the manuscript.

REFERENCES

[1] N. A. Christakis and J. H. Fowler, "Social contagion theory: examining dynamic social networks and human behavior," *Statistics in medicine*, vol. 32, no. 4, pp. 556–577, 2013.

[2] J. Coleman, E. Katz, and H. Menzel, "The diffusion of an innovation among physicians," *Sociometry*, vol. 20, no. 4, pp. 253–270, 1957.

[3] A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson, "The diffusion of microfinance," *Science*, vol. 341, no. 6144, p. 1236498, 2013.

[4] F. Liljeros, C. R. Edling, and L. A. N. Amaral, "Sexual networks: implications for the transmission of sexually transmitted infections," *Microbes and Infection*, vol. 5, no. 2, pp. 189–196, 2003.

[5] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Reviews of modern physics*, vol. 87, no. 3, p. 925, 2015.

[1] http://www.githubarchive.org/

[6] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *2010 IEEE International Conference on Data Mining*, pp. 599–608, IEEE, 2010.

[7] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks.," *ICWSM*, vol. 10, pp. 90–97, 2010.

[8] J. S. Coleman, "Social capital in the creation of human capital," *American journal of sociology*, pp. S95–S120, 1988.

[9] R. S. Burt, "The network structure of social capital," *Research in organizational behavior*, vol. 22, pp. 345–423, 2000.

[10] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, pp. 1360–1380, 1973.

[11] D. G. Rand, S. Arbesman, and N. A. Christakis, "Dynamic social networks promote cooperation in experiments with humans," *Proceedings of the National Academy of Sciences*, vol. 108, no. 48, pp. 19193–19198, 2011.

[12] I. Kawachi and L. F. Berkman, "Social ties and mental health," *Journal of Urban health*, vol. 78, no. 3, pp. 458–467, 2001.

[13] R. B. Olsen, J. Olsen, F. Gunner-Svensson, and B. Waldstrøm, "Social networks and longevity. a 14 year follow-up study among elderly in denmark," *Social science & medicine*, vol. 33, no. 10, pp. 1189–1195, 1991.

[14] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature physics*, vol. 6, no. 11, pp. 888–893, 2010.

[15] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, 2015.

[16] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, no. 6092, pp. 337–341, 2012.

[17] D. S. Massey, J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino, and J. E. Taylor, "Theories of international migration: A review and appraisal," *Population and development review*, pp. 431–466, 1993.

[18] G. S. Epstein, "Herd and network effects in migration decision-making," *Journal of Ethnic and Migration Studies*, vol. 34, no. 4, pp. 567–583, 2008.

[19] D. A. Kim, A. R. Hwong, D. Stafford, D. A. Hughes, A. J. O'Malley, J. H. Fowler, and N. A. Christakis, "Social network targeting to maximise population behaviour change: a cluster randomised controlled trial," *The Lancet*, vol. 386, no. 9989, pp. 145–153, 2015.

[20] N. A. Christakis and J. H. Fowler, "Social network sensors for early detection of contagious outbreaks," *PloS one*, vol. 5, no. 9, p. e12948, 2010.

[21] M. Garcia-Herranz, E. Moro, M. Cebrian, N. A. Christakis, and J. H. Fowler, "Using friends as sensors to detect global-scale contagious outbreaks," *PloS one*, vol. 9, no. 4, p. e92413, 2014.

[22] R. Cohen, S. Havlin, and D. Ben-Avraham, "Efficient immunization strategies for computer networks and populations," *Physical review letters*, vol. 91, no. 24, p. 247901, 2003.

[23] N. Madar, T. Kalisky, R. Cohen, D. ben Avraham, and S. Havlin, "Immunization and epidemic dynamics in complex networks," *The European physical journal b-condensed matter and complex systems*, vol. 38, no. 2, pp. 269–276, 2004.

[24] L. K. Gallos, F. Liljeros, P. Argyrakis, A. Bunde, and S. Havlin, "Improving immunization strategies," *Physical Review E*, vol. 75, no. 4, p. 045104, 2007.

[25] E. W. Zuckerman and J. T. Jost, "What makes you think you're so popular? self-evaluation maintenance and the subjective side of the" friendship paradox"," *Social Psychology Quarterly*, pp. 207–223, 2001.

[26] A.-L. Barabási, *Network Science*. Cambridge University Press, 2016.

[27] R. A. Hill and R. I. Dunbar, "Social network size in humans," *Human nature*, vol. 14, no. 1, pp. 53–72, 2003.

[28] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[29] M. E. Newman and D. J. Watts, "Renormalization group analysis of the small-world network model," *Physics Letters A*, vol. 263, no. 4, pp. 341–346, 1999.

[30] W.-X. Zhou, D. Sornette, R. A. Hill, and R. I. Dunbar, "Discrete hierarchical organization of social group sizes," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 272, no. 1561, pp. 439–444, 2005.

[31] R. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella, "The structure of online social networks mirrors those in the offline world," *Social Networks*, vol. 43, pp. 39–47, 2015.

[32] B. Fuchs, D. Sornette, and S. Thurner, "Fractal multi-level organisation of human groups in a virtual world," *Scientific reports*, vol. 4, 2014.

[33] D. d. S. Price, "A general theory of bibliometric and other cumulative advantage processes," *Journal of the American society for Information science*, vol. 27, no. 5, pp. 292–306, 1976.

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

[35] P. Holme and B. J. Kim, "Growing scale-free networks with tunable clustering," *Physical review E*, vol. 65, no. 2, p. 026107, 2002.

[36] N. Momeni and M. Rabbat, "Inferring network properties from fixed-choice design with strong and weak ties," in *Statistical Signal Processing Workshop (SSP), 2016 IEEE*, pp. 1–5, IEEE, 2016.

[37] K. Klemm and V. M. Eguiluz, "Growing scale-free networks with small-world behavior," *Physical Review E*, vol. 65, no. 5, p. 057102, 2002.

[38] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 2, 2007.

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

[40] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[41] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, vol. 42, no. 1, pp. 181–213, 2015.

[42] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, pp. 591–600, ACM, 2010.

[43] T. A. Davis and Y. Hu, "The university of florida sparse matrix collection," *ACM Transactions on Mathematical Software (TOMS)*, vol. 38, no. 1, p. 1, 2011.

[44] G. Viger, "Serendipitous recommendations for the social online collaborative network github," Master's thesis, McGill University, Montreal, QC, Canada, 2015.

[45] L. Takac and M. Zabovsky, "Data analysis in public social networks," in *International Scientific Conference and International Workshop Present Day Trends of Innovations*, pp. 1–6, 2012.

CHAPTER 7

Alter Sampling and the Friendship Paradox

7.1 Chapter Outline

In this chapter we address the question: *Why does alter sampling work?* We introduce the notion of the Friendship Paradox and its interesting social consequences. We show its connection to the effectiveness of alter sampling. We then introduce the Generalized Friendship Paradox and its instances in different contexts and discuss their implications.

7.2 Introduction

The studies discussed in the previous chapters highlight the effectiveness of alter sampling in using local information of the network structure for the efficient identification of influential nodes, to promote or detect diffusion of information or disease. After observing these practical applications, we can ask why alter sampling works. The fact that alter sampling works across diverse social networks points towards some structural property of social networks that enable this. Social networks must be organized in certain ways that directs alter sampling towards central nodes. This motivates us to

study the structure of the social networks with particular focus on the local inequalities. How do the structural properties of the ego (e.g., its centrality) relate to its network position, and to the properties of the alters? What information can we obtain from a local observation (e.g., one's degree) about the local and global organization of the network? In this chapter and the following one, we address these questions.

7.3 The Friendship Paradox

In this section we introduce the Friendship Paradox. We present its history, then we present what has been proposed in the literature as the cause of this phenomenon, and point out that those propositions are incomplete and cannot explain the observations convincingly.

Sociologist Scott Feld was first to point out (in 1991) that in friendship networks [Fel91], people have on average fewer friends than their friends do. His analysis is based on two distinct empirical friendship networks. In our terminology, Feld's argument is that if we take a sample via alter sampling and another sample via random sampling, the average degree of the former is greater.

Later, Feld's arguments were connected by Zuckerman and Jost to a widely-observed phenomenon in social psychology [ZJ01]. The phenomenon is sometimes called the 'The Lake Wobegon effect', or 'illusory superiority'. It states that most people think their abilities are higher than average. This is true even for many people who are below average [AKB+95]. This self-promoting bias is observed in different contexts. For example, people overestimate the popularity of their choices and opinions [MAC+85], think they have better memory than others [SBD99], wrongly think they perform better at tasks [KD99]. See [Hoo93] for a review. Zuckerman and Jost demonstrated that people's perception of their popularity and their standing in social networks is also subject to this self-elevating bias [ZJ01]. They gave the name *Friendship Paradox* (FP) to Feld's observations. It is a 'paradox' because most people think they are more popular than their friends, but it is not actually true.

Feld's original explanation is that the FP is merely a mathematical fluke. He argues that the FP is similar to the "class size paradox" [FG77]. Suppose a school has four classes, three of them with capacity 10 and one of them with capacity 70. The average class size is $\frac{3\times10+7}{4} = 25$. But suppose to calculate average class size, we

ask every student in the school to report their class size, and then take the average of the reported values. We will get 70 students reporting 70, and $3 \times 10 = 30$ students reporting 10. If we take the average of these reported values, we obtain $\frac{70\times70+30\times10}{70+3\times10} = 52$, which is larger than the actual value. Feld argued that the FP is essentially the same phenomenon. Let us denote the degree of individual x by k_x, the average degree of the network by $\overline{k}$, the average degree of friends by $\overline{k}_{\rm f}$, and the number of individuals by N. Feld argues that $\overline{k}_{\rm f}$ is calculated as follows: survey all individuals and ask them about the degree of their friends. Individual x reports k_x entries, each being the degree of a friend. So the number of total entries we collect is $\sum_x k_x$. We denote the frequency of entry x by w_x, which means w_x people mentioned degree of individual x. The average of the entries is

$$\frac{\sum_x w_x k_x}{\sum_x k_x} = \frac{\sum_x w_x k_x}{N\overline{k}} = \frac{(\sum_x w_x k_x)/N}{\overline{k}}. \tag{7.1}$$

It is easy to calculate w_y. In this list of entries, individual x appears k_x times, being mentioned by each of its friends once. So $w_x = k_x$ If we take the average of these entries, the result is $\overline{k}_{\rm f} = \frac{(\sum_x k_x^2)/N}{\overline{k}}$. but we have

$$\overline{k}_{\rm f} = \frac{(\sum_x k_x^2)/N}{\overline{k}} = \frac{{\rm var}(k) + \overline{k}^2}{\overline{k}} = \frac{{\rm var}(k)}{\overline{k}} + \overline{k}, \tag{7.2}$$

so we have

$$\overline{k}_{\rm f} - \overline{k} = \frac{{\rm var}(k)}{\overline{k}} > 0. \tag{7.3}$$

From this Feld concluded that $\overline{k}_{\rm f}$ is trivially greater than $\overline{k}$, and that it is just an effect of binning.

In Chapter 8 we show that this can not explain the phenomenon fully. Because, as we showed in Chapter 6, the vast majority of nodes in social networks have degrees smaller than the mean *and* the median of their neighbors. While the above mathematical explanation would necessary hold for any network, the median version of the FP would not. A network must be organized in certain ways so that the median FP would also occur. We demonstrate in Chapter 8 that this sheds light on an interesting and important feature how social networks are organized. Note that the FP can also be viewed as a local inequality: ego being less central than the alters. How can we measure

and quantify the FP? Do all the nodes in the network experience the FP? What are the consequences of high or low prevalence of the FP on the organization of the social network? Chapter 7, presents our contribution which addresses these questions. We classify the FP into different types, introduce measures for quantifying it, study its prevalence on an online social network, and discuss its implications about the way the network are organized.

It is by now clear why alter sampling and the FP are closely related concepts. Alter sampling works because alters are more well-connected than the ego, which is what the FP states. If social networks were organized differently, alter sampling would not prove very effective. For example, consider an Erdős-Rényi random network with size N and link probability p. Define a random variable L_j for link j, which is 1 if the link exists and is 0 otherwise. It is a Bernouli random variable with probability p, by definition. The average degree of the network is twice the number of links divided by N, that is, twice the sum over all L_j values (sum over links), divided by N. There are $\frac{N(N-1)}{2}$ links. The expected value of the sum of Bernouli random variables is $\frac{N(N-1)}{2}p$. Thus the expected average degree is $\overline{k} = (N-1)p$. Now we find $\overline{k}_{\mathrm{f}}$ using Equation (7.2). The degree distribution of an Erdős-Rényi random network is $p(k) = \binom{N-1}{k} p^k q^{N-1-k}$, where $q \stackrel{\text{def}}{=} 1-p$. To obtain the variance, we have

$$\mathrm{var}(k) = \left[\sum_k k^2 p(k)\right] - \left[(N-1)p\right]^2 = \left[\sum_k k^2 \binom{N-1}{k} p^k q^{N-1-k}\right] - \left[(N-1)p\right]^2. \tag{7.4}$$

First we perform the following summation for arbitrary x and y and integer M:

$$\begin{aligned}
F(x,y,M) &= \sum_{k=0}^{M} k^2 \binom{M}{k} x^k y^{M-k} = \sum_{k=0}^{M} \left[k(k-1)+k\right] \binom{M}{k} x^k y^{M-k} \\
&= \left[\sum_{k=0}^{M} k(k-1) \binom{M}{k} x^k y^{M-k}\right] + \left[\sum_{k=0}^{M} k \binom{M}{k} x^k y^{M-k}\right] \\
&= \left[x^2 \frac{\partial^2}{\partial x^2} \sum_{k=0}^{M} \binom{M}{k} x^k y^{M-k}\right] + \left[x \frac{\partial}{\partial x} \sum_{k=0}^{M} \binom{M}{k} x^k y^{M-k}\right] \\
&= \left[x^2 \frac{\partial^2}{\partial x^2} (x+y)^M\right] + \left[x \frac{\partial}{\partial x} (x+y)^M\right] \\
&= x^2 M(M-1)(x+y)^{M-2} + xM(x+y)^{M-1}.
\end{aligned} \tag{7.5}$$

Using this result, from (7.4) we have

$$\frac{\mathrm{var}(k)}{\overline{k}} = \frac{\Big[F(p,1-p,N-1)\Big] - \Big[(N-1)p\Big]^2}{(N-1)p}$$
$$= \frac{\Big[p^2(N-1)(N-2) + p(N-1)\Big] - \Big[(N-1)p\Big]^2}{(N-1)p}. \quad (7.6)$$

Inserting this into (7.2), we get

$$\overline{k}_{\mathrm{f}} = \frac{\mathrm{var}(k)}{\overline{k}} + \overline{k} = p(N-2) + 1 - p(N-1) + p(N-1) = (N-1)p + (1-p) \quad (7.7)$$

This means that the expected gain of alter sampling is

$$\frac{\overline{k}_{\mathrm{f}}}{\overline{k}} = 1 + \frac{1-p}{\overline{k}} < 2. \quad (7.8)$$

Our empirical observations in Chapter 6 showed that the disparity is much greater, sometimes even up to two orders of magnitudes. So alter sampling does not necessarily provide a large gain for every network structure. Social networks must have special organizations to enable its large gain.

In the above discussions, the definition of the FP was based on the degrees of nodes. That is, the property of the individuals that is being compared to their neighbors was their degrees. We now expand this notion and show that a similar phenomenon exists for personal attributes, that is, non-network characteristics of individuals. To that aim, we first highlight why such non-network properties matter in research on networks at all.

7.4 Introducing the Generalized Friendship Paradox

The units of social networks are people. They have properties beyond network statistics; individuals have personal attributes as well. These attributes can interact with the structure. We briefly discuss how personal attributes and network position mutually interact. That is, why should a network researcher care about personal attributes of nodes.

The position of the ego in the social network affects its social life. It determines the information and resources the ego has access to. It also determines with whom the ego can interact and share. These availabilities can affect the social attributes and attitudes of the ego, e.g., how social or isolated the ego is. But the relation is bidirectional. The position of the ego in the social network is itself affected by the attributes and attitudes of the ego. Attributes and attitudes are simultaneously products and determinants of network position. For instance, in friendship networks, people are not equally social and outgoing. This affects their social network, which in turn affects their opportunities for association. In scientific collaboration networks, scholars have different levels of productivity and activity. These traits affect their attractiveness to peers for collaborations, and shape their collaboration network. This structure in turn influences their chances of future collaborations. In online social networks, users have heterogeneous levels of content generation and sharing. Some users generate more interesting or important content, or are quicker than others in detecting and spreading such content. These traits affect one's network (e.g., how many followers one gets on Twitter, and who those followers are). The structure of this network will in turn affect the content users receive and the chances they have in spreading their posts. In all these cases, we see that the personal attributes of individuals can interact with their position in the network.

Furthermore, some attributes are personal (such as being funny in social interactions, or creativity and activity in content generation in social networks). These attributes pertain to ego's influence on its own behavior (which may or may not influence others' behavior via interaction). On the other hand, some attributes are interpersonal (such as peer influence and charisma). Their definition is based on the influence of the ego over the behavior of alters. This further emphasizes the importance of personal attributes on the structure of the social network. This is interesting because by looking at the structure of a network, which is just a structure of the social group, we can say things about the personal traits, and conversely, we can look at the effects of network structure on the individual. For example, economic experiments have shown that people who are more cooperative tend to receive more links from others because others would like to increase their chances of constructive mutually-beneficial relationship [RAC11]. Conversely, sociological studies have shown that dense high-clustering

social structure provides high social support, and at the same time, high social control and norm enforcement [Col88].

In short, an interesting feature of the field of social networks is that it relates individual properties to structural properties *and* the converse. So similar to studying the effects of structure on the individual, the converse effects are also important to study.

If we can study the structural inequalities of social networks by comparing the degrees of nodes to those of their neighbors, we can also do this for non-network individual attributes. That is, we can study the relation between the personal attributes of the neighbors in the network. Similar to the case of degrees, one can study the local inequalities for nodal attributes. So the notion of the FP can be extended to attributes. This results in the *Generalized Friendship Paradox* (GFP). For example, it has been shown that in scientific collaboration networks, scholars on average have fewer citations than their collaborators [EJ14]. Same is true for the H-index [BLA15][1] .

One might assume the following explanation for this phenomenon: degree (in this case, number of collaborators) and personal attributes (e.g., the H-index) are positively correlated, and this means that whenever the FP exists, GFP would also exist [EJ14]. In Chapter 8, we show that this intuitive explanation is incomplete. We present a case study on a large network data set, we measure several nodal attributes that have social impacts, study their interrelation with network structure, and study the local inequalities with regard to these nodal attributes. We show that the GFP is highly prevalent even for attributes who do not have a significant correlation with degree.

7.5 Chapter Summary

We began with highlighting the effectiveness of alter sampling across diverse social networks and related it to the Friendship Paradox, which is a property that results from how social networks are locally organized. The FP is an interesting phenomenon relating to local structural inequalities of social networks. We then argued that individual

[1] The H-index of a scholar is defined as the minimum k such that the scholar has no less than k papers with no less than k citations.

attributes are also important and can affect the structure of social networks just as the converse is true. This highlights importance of local inequalities in attributes, which is the focus of the paper presented in the next chapter.

CHAPTER 8

Paper: Friendship Paradox and the Inequalities in Social Networks

The material presented in this chapter was published in the following journal:

N. Momeni, M. Rabbat, "Qualities and Inequalities in Online Social Networks through the Lens of the Generalized Friendship Paradox", PloS one 11.2 (2016): e0143633.

Please note that the references of the manuscript are listed at the end of this chapter.

Qualities and Inequalities in Online Social Networks through the Lens of the Generalized Friendship Paradox

Naghmeh Momeni and Michael G. Rabbat

Abstract

The friendship paradox is the phenomenon that in social networks, people on average have fewer friends than their friends do. The generalized friendship paradox is an extension to attributes other than the number of friends. The friendship paradox and its generalized version have gathered recent attention due to the information they provide about network structure and local inequalities. In this paper, we propose several measures of nodal qualities which capture different aspects of their activities and influence in online social networks. Using these measures we analyze the prevalence of the generalized friendship paradox over Twitter and we report high levels of prevalence (up to over 90% of nodes). We contend that this prevalence of the friendship paradox and its generalized version arise because of the hierarchical nature of the connections in the network. This hierarchy is nested as opposed to being star-like. We conclude that these paradoxes are collective phenomena not created merely by a minority of well-connected or high-attribute nodes. Previous papers had argued that positive correlation between degrees and attributes results in the generalized friendship paradox. Our results show that although such a positive correlation is sufficient for the generalized friendship paradox, it is not necessary.

INTRODUCTION

The *friendship paradox* (FP), first introduced by Feld [1], is a phenomenon stemming from the structural properties of social networks. It indicates that although most people think that they are more popular than their friends, in actuality the converse is true: on average, each person has fewer friends than his/her friends do. This observation sheds light

on the local inequalities of social networks, how people organize their social ties, and how these inequalities extend to macro structures of social networks. This paradox has also been observed in online settings as well [2], [3]. It has been contended that this paradox can be exploited for early detection of flu outbreaks [4], [5], and more generally, finding well-connected nodes in large networks [6]–[8].

The generalized version of the friendship paradox is an extension to attributes other than number of social ties. It was introduced in [6], where it is shown that in scientific collaboration networks, each scholar has on average fewer citations than his/her collaborators do. The *generalized friendship paradox* (GFP) has been studied analytically in [9], [10]. The GFP links intra-personal attributes to inter-personal ties, and thus sheds light on the interplay between nodal characteristics and network structure. It also takes a notable step towards characterizing the local inequalities of networks regarding non-structural nodal properties. In this paper, we study the GFP in the context of online social networks, and we consider nodal attributes that corresponds to *influence*. We study how the structure of connections between nodes is related to the influence they have upon others.

Finding highly-influential nodes in social networks is useful for a variety of tasks such as understanding diffusion of information [11]–[13] and misinformation [14], [15], promoting cooperation [16], [17], optimal product placement for marketing purposes [18], [19], optimal immunization and vaccination strategies [20], and studying the diffusion of innovation [21].

Characterizing influence is not straightforward. Being highly connected does not necessarily mean being influential. For example, in online social media, it has been found that highly-connected individuals are overwhelmed by information flows and sometimes cannot detect viral content effectively [22]. Thus, purely-structural measures alone (such as degree) cannot capture nodal influence. Furthermore, there are different types of influence, and nodes with different patterns of activities and impacts can be deemed influential. Most of the online social networks include initiation and adoption processes. Each user can generate contents visible to other users who can re-post them. We need measures of influence that enable us to compare, for example, highly-active nodes with average clout with occasionally-active nodes with great clout.

This paper proposes measures that capture multiple aspects of node activity and influence in online social networks. Two of these measures quantify nodal activity, and four of them quantify inter-nodal influence. We compute and analyse these measures on Twitter (the micro-blogging platform), but the measures are general and as we contend, they can be applied to any network setting with initiation and adoption mechanisms. We study the distributions and statistical properties of these measures.

In this paper, we study the GFP (on the *individual level*, in the terminology of [6], [9]) through the lens of the measures of activity and influence that we introduce. We also introduce new measures for quantifying the GFP. Our measures assess to what extent nodes of a network experience the GFP. Throughout the paper, we use the term 'Neighbor Superiority' [23] to refer to this phenomenon, because we found no evidence in the literature that, for example, most scholars think that they are more cited than their collaborators (which would contradict reality and create a 'paradox'). Hence, the word paradox is not appropriate for contexts other than friendship. Furthermore, in online social media, evidence point towards the opposite direction [24], and most users assess their neighbors more highly than themselves.

We find high prevalence of neighbor superiority both in terms of connectivity and quality. For each of these nodal attributes, a vast majority of the nodes, even those who rank very highly in the population (for example, among the top 0.5% in terms of tweeting activity, or in terms of popularity), experience neighbor superiority. We analyse the distributions of the measures and their prevalences more closely and uncover a hierarchical nature in the connectivity of the Twitter graph. We contend that neighbor superiority, and its special case, the friendship paradox, are not mere mathematical artifacts that result from a star-like structure—a simplistic picture in which almost all nodes are connected to a few hubs, and these hubs make them experience neighbor superiority [1]. Instead, we show that there is a hierarchical nature in the pattern of connections in the network. Moreover, similar to the friendship paradox that enables biased sampling and detection of popular nodes [5], [6], our results indicate that the same scheme can be applied in terms of non-structural nodal attributes (i.e., qualities), to detect high-quality nodes.

The rest of this paper is organized as follows. We first introduce the terminology used throughout the paper. After describing the data set, we introduce measures of nodal quality (activity and influence), as well as measures to quantify neighbor superiority. We then present the distribution of nodal qualities and connectivity. We discuss results on neighbor superiority and focus on their implications for the underlying structure of the Twitter graph.

TERMINOLOGY

On Twitter, users can post short texts that should not exceed 140 characters. These posts are called *tweets*. User can post *original tweets*, or can repost another user's tweet, which is called *retweeting*. Each user can *follow* other users. When user A follows user B, we say that A is a *follower* of B and B is a *followee* of A. When A follows B, A subscribes to the tweets posted by B. Each user can see the tweets of his/her followees on his/her home Twitter feed.

The underlying web of connectivities between users can be modelled as a graph. Users are mapped onto *nodes*, and their connections onto *links*. Note that on Twitter, B can follow A back or not. Mathematically, this means that the Twitter graph is *directed*. There are two types of adjacency relationships that can be defined on the Twitter graph—follower and followee. We use the term *neighbor* to refer to both of these types of connection. So for each user, an neighbor can be either a follower or a followee. The number of followers and the number of followees of a user are called the *in-degree* and *out-degree* of that user, respectively.

When user A posts a tweet, the followers of A can see it. When one of its followers, say B, retweets it, the followers of B can also see the tweet. With each retweet, the number of users who are exposed to the tweet increases. This is called a *cascade* of the original tweet which was posted by A. The total number of retweets that a tweet by A receives is called the *cascade size* of that tweet. Note that this retweet can be done either by A's own followers, or the followers of A's followers, and so on. After user B retweets one of A's tweets, then if a user C, which is a follower of B retweets that tweet, this retweet is counted only for the original tweet, and only for A. In other words, any cascade only

has one root tweet and one initiating user. All the retweets are counted for that tweet and that user. This mechanism is internal to Twitter and we follow the same convention in this paper.

There are two categories of attributes that we consider in this paper. The first category consists of the in-degree and out-degree, which are *structural* attributes. The second category assesses tweeting activities of a node. We denote the attributes that belong to the latter category by *quality* (Note that by *quality* we mean an intrinsic fitness value that drives the connection and following patterns. It does not signify any quality of the content of the tweets). We define 6 different qualities in the section .

DATA

We use two datasets in this paper. The first one is presented by Yang and Leskovec [25] and contains over 470 million tweets by over 18 million users, which capture over 20% of all tweets posted over a 7-month period, starting from June 2009. For the network of connectivity of Twitter users, we use the dataset collected by Kwak et al. [26]. This dataset comprises all the links between users who joined Twitter prior to August 2009. We only consider users that are present in both data sets. The subgraph of connectivity has 5.8 million users and over 193 million links. The subset of all tweets that is considered includes over 200 million tweets.

METHODS

We discard repeated tweets for each user and count only the number of distinct tweets. Following the convention mentioned above, we count unique retweets only for the root user of a cascade, not for other users who retweet the message (who consequently received further retweets for their retweet). So the retweets for each user can be from those who directly follow the user, or the followers of the followers of the user, and so on.

Node Attributes as Quality

Minding the specificities of the data set at hand, we considered six possible candidates as measures for node qualities. Two of these measures quantify the activity of nodes, and four

of them quantify their influence on others. Combining these six measures with in-degree and out-degree, we construct an 8-dimensional feature vector that characterizes each node. The six quality features are the following:

1) The *number of tweets* (NT) is the total number of posts of the user, which includes original tweets and retweets.
2) The *number of original tweets* (NOT) is the number of tweets that the user has initiated.
3) The *total times retweeted* (TTR) is the number of times that the posts initiated by the user got retweeted by other users. It is the gross number of retweets that the user has received.
4) The *number of tweets retweeted* (NTR) is the number of tweets initiated by the user that received at least one retweet from other users. In other words, the NTR is the number of times that the user has created a cascade.
5) The *retweets per tweet* (RPT) is the average number of retweets received by a tweet initiated by the user. In other words, the RPT of a user is the expected cascade size that the user engenders.
6) The *fraction of tweets retweeted* (FTR) is the normalized version of NTR, that is, it is equal to the fraction of tweets initiated by the user that received at least one retweet from other users. This characterizes the clout of the user by assessing the likelihood that a tweet initiated by the user will engender a cascade.

As mentioned above, there are two categories: activity and influence. NT and NOT are measures of activity: NT is a measure of *total activity*; it measures how much a user posts tweets (that can be created by him/her or his/her peers). NOT is a measure of *total novel activity*.

The next four measures (TTR, NTR, RPT, and FTR) are measures of influence: TTR measures the *total influence* of a node over followers. NTR is a measure of *total success*; it measures the number of successful initiations. RPT is a measure of *efficiency*; it measures the expected influence per initiation. FTR is a measure of *consistency*, which measures the likelihood of generating cascade of any size per initiation.

Note that these measures are general, and need not be confined to Twitter. NT and NOT can be used for any social network in which a specific action can be defined as 'activity' (e.g., posting content on Google+, Pinterst, Instagram, Tumblr, etc.). TTR, NTR, RPT and FTR can be used in any network context in which there is a mechanism for sharing, reposting, or adoption. For example, on Pinterest, users can *pin* items on their boards, and their followers can *re-pin* them (equivalent of retweeting). On Facebook, posts can be *shared* and on Tumblr, users can *re-blog* the posts by other users.

Through a hypothetical example, we shed light on the nuances of these measures and the different aspects of the users they capture. Let us consider three users: user 1 has made 100 original tweets, one of them has received 1000 retweets and the rest have received none. User 2 has posted 100 original tweets; each of them have received 10 retweets. So in total, user 2 has received 1000 retweets. User 3 has made only 10 tweets;, each of them have received 50 retweets. So user 3 has received 500 retweets in total.

Users 1 and 2 have equal TTR values of 1000, which exceeds that of user 3 (which is 500). Note that TTR cannot distinguish between the first two users, while their patterns of influence are clearly different. We can distinguish between user 1 and 2 using FTR, because the FTR of user 1 is 0.01, whereas the FTR of user 2 is equal to 1. In this example user 1 has had a moment in the sun, and does not have a steady influence over other users, whereas user 2 consistently creates cascades (of smaller size as compared to that of user 1). In other words, user 2 is more reliable to engender cascades than user 1, but the cascade is not as large.

Now let us consider users 2 and 3. For both of them, the FTR is 1, which means that for both users, every tweet has has been retweeted at least once. However, the RPT of user 3 is 50, and the RPT of user 2 is 10. This means that although user 3 is not as active as user 2, the cascades created by user 3 are on average five times larger.

We now turn our attention to the distribution of these different measures of quality in the network. In addition to the inequalities of the degrees, we investigate the inequalities between node qualities that exist in the network.

Measures of Neighbor Superiority

In this section we introduce measures to quantify neighbor superiority, which is the essence of the friendship paradox and its generalized version. These measures are generalizations of the measures we introduced in [27], which pertain to undirected networks only.

In the present paper, the network under consideration is directed. However, to develop intuition about the measures that we are going to introduce, first let us consider a simple undirected network. So there is no follower/followee distinction; rather, each node simply has neighbors. In this case, we can compare the degree of each node with, say, the average of the degrees of its neighbors. If the degree of the node is smaller than the average of its neighbors' degree, we say that the node is experiencing mean neighbor superiority. Throughout the network, different nodes with different degrees can be experiencing mean neighbor superiority. A question we can ask about the network under study is that, how large should the degree of a node be so that it will not experience mean neighbor superiority? To address this question, we introduce the notion of *critical degree for the mean*, which is defined to be the maximum of the degrees of all the nodes in the network that experience mean neighbor superiority. In other words, no node in the network with degree greater than the critical degree experiences mean neighbor superiority. Let us denote the set of neighbors of node x by N_x, and let us denote the degree of node x by k_x. The critical degree for the mean can be expressed as follows:

$$\widetilde{K} = \max \left\{ k_x \middle| k_x < \frac{\sum_{y \in N_x} k_y}{|N_x|} \right\}. \tag{1}$$

A drawback of using the mean neighbor degree is that a node might be experiencing neighbor superiority only because one of its many neighbors had a very large degree, hence making the mean neighbor degree large. We can also compare the degree of each node with the median neighbor degree, which alleviates the problem of outliers (see [27], [28]). A node is said to be experiencing median neighbor superiority if more than half of its neighbors have higher degrees than it does. Let us denote the median by $M(\cdot)$. Similar to

the mean, we can also define the *critical degree for the median* as follows:

$$\widehat{K} = \max\left\{k_x \middle| k_x < M\left(k_y | y \in N_x\right)\right\}. \quad (2)$$

Note that any nodal attribute can be compared to the mean or median of the neighbors in order to define neighbor superiority. It need not be degree. It can be age, for example.

In addition to critical values, another way of quantifying neighbor superiority in the network would be to measure the prevalence of neighbor superiority, that is the fraction of nodes who experience a given type of neighbor superiority. For example, we can ask what fraction of nodes experience median neighbor superiority or mean neighbor superiority.

Now let us extend these definitions to the case of a directed network. In this case, each node has two distinct sets of neighbors: followers and followees. Thus, we can compare each attribute of a node to its followers and its followees. The results of these comparisons need not be the same (in fact, as we will demonstrate, they are not). This proliferates the number of ways we can compare a node to its neighbors. For example, we can compare the in-degree of a node (i.e., number of followers) with the in-degree of its followers, or the in-degree of its followees. We can also use measures of quality, as introduced above. For the same reasoning as mentioned for the undirected case, we can use both the mean and the median for comparison. This engenders several possible ways of defining neighbor superiority, as well as corresponding critical values (both for the mean and median versions).

In total, there are 32 different critical values that can be defined: median/mean follower/followee superiority for 8 different possible attributes. Corresponding to these 32 different types of superiorities, we can also measure 32 fractions that reflect what fraction of nodes in the network experience a given type of superiority.

RESULTS AND DISCUSSION

Distribution of Quality and Degree

We computed the 6 measures of quality, as well as in-degree and out-degree, for all the nodes in the Twitter network. The distribution of all the 8 nodal attributes are highly skewed. Table I presents the summary statistics of the distributions. The percentage headers

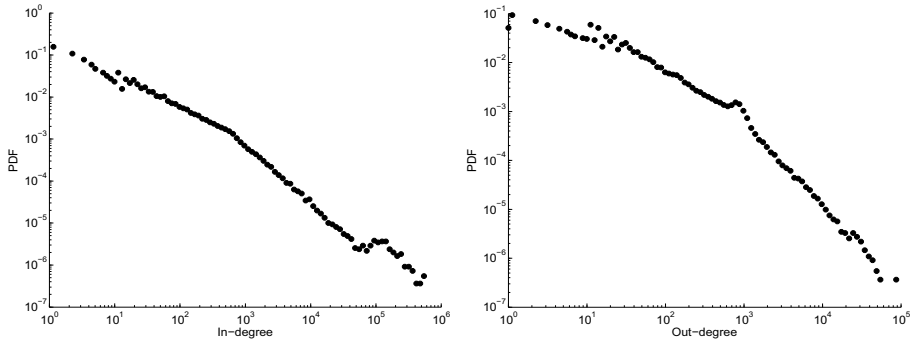
indicate percentiles: e.g., the first row of the 90% column is 41, which means that 90% of the users have in-degree less than or equal to 41. It is of note that the four zeros at the bottom of the third column indicate that over 75% of the users never got retweeted, which implies a high skew in the distribution. In other words, on Twitter, most people only observe. They read, but seldom retweet.

	Mean	Median (50%)	75%	90%	95%	99%	Max (100%)
In-Degree	35.2	4	13	41	92	459	625520
Out-Degree	35.2	10	22	57	115	539	86800
NT	37.3	5	22	79	157	526	85316
NOT	35.1	4	20	74	148	499	85234
TTR	2.40	0	0	2	6	35	82036
NTR	0.83	0	0	1	3	13	4803
RPT	0.167	0	0	0.056	0.167	1.224	12567.0
FTR	0.02	0	0	0.036	0.100	0.500	1.0

TABLE I: Summary statistics of nodal attributes.

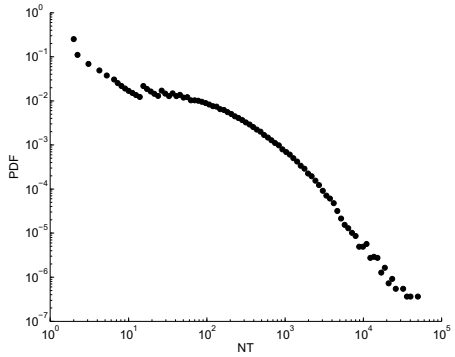
The distribution of the 8 nodal attributes is depicted in Figure 1. For each attribute, we divide the interval between the minimum and maximum values of the attribute into 100 bins. The bin sizes increase logarithmically and the distributions are plotted on a log-log scale. Note that for each bin the logarithm of the endpoint is evaluated. All nodal attributes exhibit a heavy-tailed distribution. Figures 1g and 1h exhibit interesting behavior: the majority of the nodes (over 80%) have zero RPT and zero FTR. Setting these users aside, the rest of the population exhibit distributions for RPT and FTR that, unlike other 6 nodal attributes, are not monotonically decreasing.

Table II presents the correlation coefficients between the six measures of influence, as well as in-degree and out-degree (28 pairs in total). It can be observed that the the magnitude of correlation between 21 pairs are below 0.15, and only one is above 0.5. This confirms that these measures capture distinct components of nodal attributes. Also note that in-degree and out-degree are not highly correlated with measures of influence. This is a notable observation, and further confirms that measures of connectivity are not

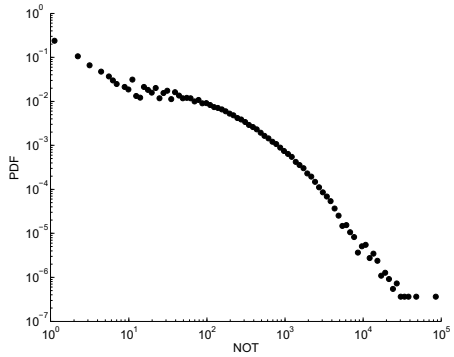


(a) In-degree

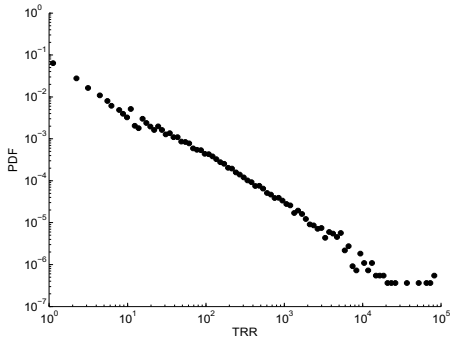
(b) Out-degree



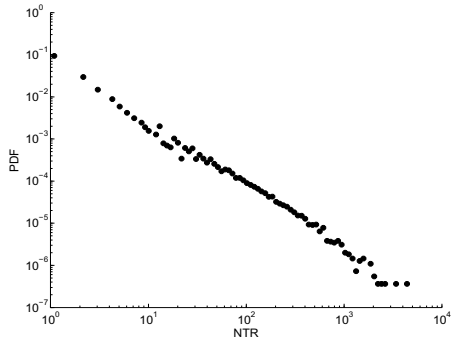
(c) NT: Number of Tweets



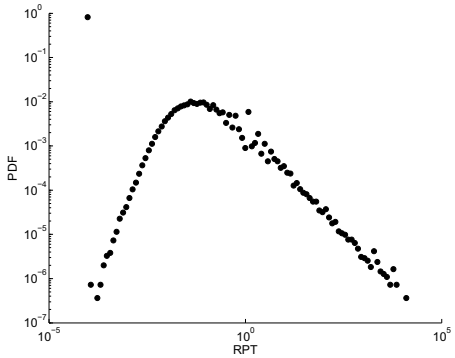
(d) NOT: Number of Original Tweets.



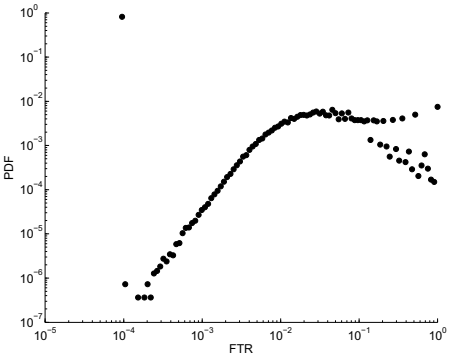
(e) TTR: Total Times Retweeted.



(f) NTR: Number of Tweets Retweeted.



(g) RPT: Retweets Per Tweet.



(h) FTR: Fraction of Tweets Retweeted.

Figure 1. Distributions of different nodal attributes.

	In	Out	NT	NOT	TTR	NTR	RPT	FTR
In	1	0.271	0.041	0.040	0.067	0.112	0.002	0.002
Out	0.271	1	0.128	0.121	0.042	0.156	0.002	0.053
NT	0.041	0.128	1	0.993	0.108	0.486	-0.001	0.013
NOT	0.040	0.121	0.993	1	0.086	0.419	-0.002	-0.002
TTR	0.067	0.042	0.108	0.086	1	0.231	0.356	0.054
NTR	0.112	0.156	0.486	0.419	0.231	1	0.003	0.112
RPT	0.002	0.002	-0.001	-0.002	0.356	0.003	1	0.075
FTR	0.002	0.053	0.013	0.002	0.054	0.112	0.075	1

TABLE II: Correlation coefficients between nodal attributes

necessarily correlated with influence. Had the correlation coefficient between degrees and nodal attributes been large, then the GFP would become an artifact of the FP. However, this is not the case, and they cannot be ascribed to a common cause. Finally, note that NT and NOT are highly correlated, which is expected. We have included both because in the calculation of other quantities NOT was employed.

Results on Neighbor Superiority

The fraction of nodes in the network that experience the corresponding types of neighbor superiority are presented in Table III. The critical values pertaining to each type of neighbor superiority is presented in Table IV.

A hierarchy of connections can be discerned from Table III. For any given type of neighbor superiority, either in the mean or the median version, we observe that the fraction of nodes experiencing followee superiority exceeds the fraction of nodes experiencing follower superiority (which is true in 15 out of all 16 possible cases). This suggests the existence of a hierarchy of attachment, which is a result of the tendency of users to follow those who have higher attributes than them, both in terms of degree (in/out) and quality.

For all types of mean and median followee superiority the fractions of nodes experiencing the corresponding neighbor superiority are above 80% and 66%, respectively. Table III shows that the fraction of nodes experiencing 12 out of 16 types of median superiority

		Fraction of nodes experiencing neighbor superiority (%)			
		Mean		Median	
		Follower	Followee	Follower	Followee
Structural	In-degree	85.5	93.7	79.7	90.2
	Out-degree	86.1	92.5	82.0	80.7
Quality: Activity	NT	71.4	87.2	58.4	79.3
	NOT	71.2	87.2	57.8	79.4
Quality: Influence	TTR	65.9	83.3	33.0	67.8
	NTR	65.2	83.1	32.5	67.2
	RPT	64.4	81.9	34.2	66.9
	FTR	63.0	80.4	34.0	66.5

TABLE III: Fraction of nodes experiencing different types of neighbor superiority.

		Critical Values			
		Mean		Median	
		Follower	Followee	Follower	Followee
Structural	In-degree	2890	155657	1894	114629
	Out-degree	2887	2887	2108	1997
Quality: Activity	NT	3305	5009	1837	5009
	NOT	2853	5009	1827	5009
Quality: Influence	TTR	540	2590	628	1962
	NTR	219	301	141	286
	RPT	54.1	64.5	40.0	19.0
	FTR	0.975	0.911	0.975	0.896

TABLE IV: Critical values for different types of neighbor superiority.

is higher than 57%. This means that for these users, more than half of the users they are connected to have higher attributes than them. This challenges the simplistic picture that reduces neighbor superiority to a mere statistical artifact. This simplistic picture contends that neighbor superiority merely results from the existence of a few well-connected nodes with high attributes that make their neighbors experience mean neighbor superiority by lifting their neighbor-averaged attributes. We observe that for most of the nodes, it is not a single dominant neighbor that makes them experience neighbor superiority; rather, it is more than half of their neighbors that do this collectively.

The values of bottom half of the column pertaining to median follower superiority are smaller, as compared to other figures in Table III. Note that these four correspond to the four zeros in the 75% percentile column of Table I. We can explain this observation saying that for a large majority of the network (over 75%), the values of TTR, NTR, RPT and FTR are equal to zero. It is plausible to deduce that these values are also zero for the majority of the followers of each of these users. This renders the median value of their followers equal to zero, which makes them not experience follower superiority. However, since even one nonzero follower suffices to lift the mean above zero, the fraction of nodes experiencing mean follower superiority is much larger than those experiencing median follower superiority (the range of fractions for the mean version is between 63% and 66%, whereas for the median version, the range is between 32% and 34%). Comparing these fractions with the corresponding figures on the rightmost column of Table III provides further evidence for the existence of a following hierarchy. In short, users rarely follow down, they mostly tend to follow up or across.

The critical values also provide insight into the hierarchical structure of the connectivity of the Twitter network. For example, for the mean followee superiority in in-degree, the critical value of 155657. This means that even a user who has 155657 followers follows users who on average have more followers than him/her. Noting that only 1% of the users have more than 460 followers (Table I), the user with 155657 ranks very highly in terms of number of followers (99.99 percentile), and even this user is following those who on average have more followers. Similarly, for the median followee superiority in in-degree,

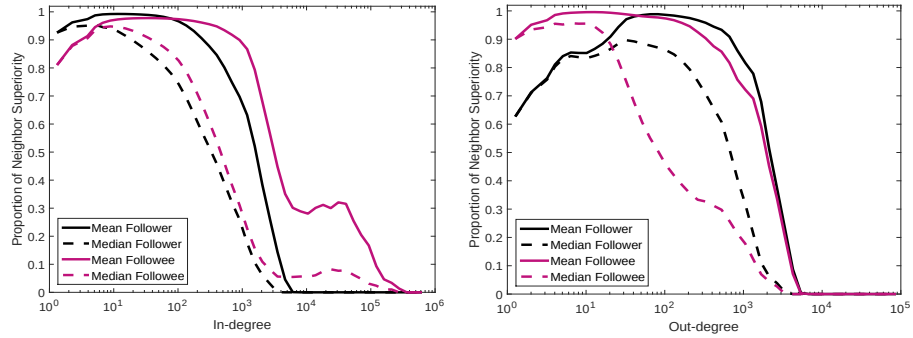
the critical value is 114629. This means that even for a user with this many followers, the majority of his/her followees have more followers than s/he does.

In addition to degrees, similar observations can be made on nodal qualities. Consider TTR as an example. For the mean followee superiority in TTR, the critical value is 2590. Note that, as Table I presents, more than 75% of all the users never get retweeted, that is, they have TTR of zero. Even a user with such a high value of TTR experiences mean followee superiority. Similarly, for RPT, Table I tells us that 95% of all the users have RPT values below 0.16. From Table IV we observe that the critical values for the mean followee superiority in RPT is 64.5. Even a user with RPT as large as 64.5 follows users that are on average more influential in terms of cascade size. These provide evidence for the hierarchical nature of the connections of the Twitter network both in terms of connectivity and in terms of nodal qualities.

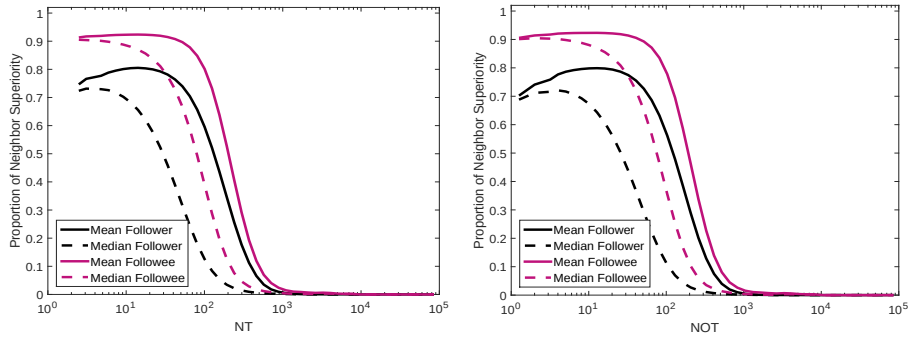
Figure 2 illustrates the empirical distribution of experiencing superiority pertaining to different attributes. The horizontal axis is log-scaled for better visibility. We construct 50 bins in each case, and for all the users who fall in each bin, we calculated the fraction of them who experience the given type of superiority.

The intuitive expectation for any type of superiority might be that the higher the attribute of a node is, the less likely it would be for that node to experience neighbor superiority. In other words, one might expect to observe a uniformly decreasing likelihood of experiencing neighbor superiority as a function of any nodal attribute, which is maximized when the value of the attribute takes its minimum values. However, in Figure 2, only two subfigures resemble this scheme, which are Figures 2e and 2f. All other subfigures present either curves with plateaus, or unimodal ones. This is a crucial observation with interesting consequences, as we shall discuss in detail next. Note that the point where the curves hit the x-axis are pertained to the corresponding critical values.

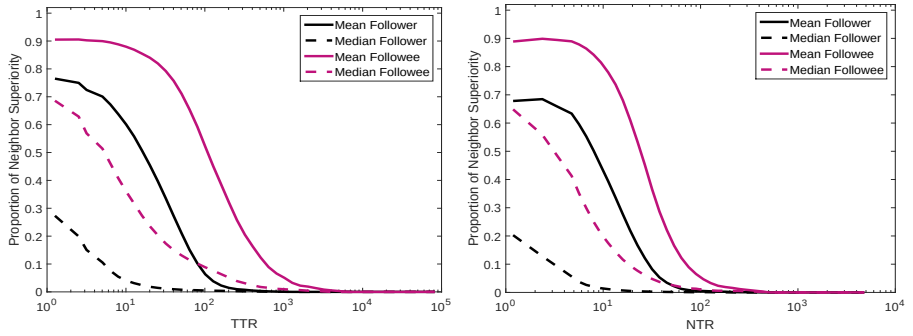
Let us consider Figure 2a as an example, and let us first focus on the solid red curve. For the minimum degree, we observe that that proportion of experiencing mean followee superiority is 0.8, which means that 20% of the users with minimum in-degree do not experience mean followee superiority. The proportion *increases* to above 0.95 when the



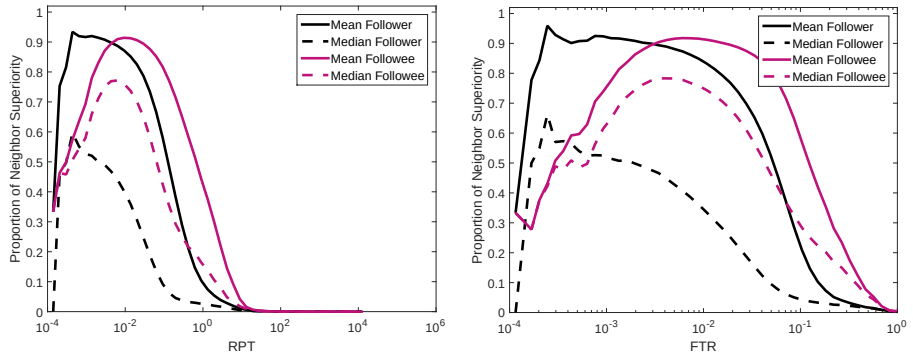
(a) In-degree **(b)** Out-degree



(c) NT: Number of Tweets **(d)** NOT: Number of Original Tweets.



(e) TTR: Total Times Retweeted. **(f)** NTR: Number of Tweets Retweeted.



(g) RPT: Retweets Per Tweet. **(h)** FTR: Fraction of Tweets Retweeted.

Figure 2. Proportion of neighbor superiority of different types for different nodal attributes.

degree is 5, then follows a wide plateau up to degree of 1000, and then decreases. The proportion of experiencing followee superiority is 0.3 for an in-degree as high as 20000: of the users with around 20000 followers, 30% follow users with, on average, more followers. So critical values do not pertain to outliers. In other words, a high critical value cannot be construed as "maybe this is the only node among all nodes with high values of that attribute that experiences neighbor superiority". Rather, a high critical value suggests a continuous decline in the proportion of experiencing neighbor superiority. A high critical value does suggest that in the network under consideration, even the nodes with high values of the attribute experience neighbor superiority.

Note that Figure 2a is consistent with Figure 3c in [9], where an analytical approach is undertaken and the proportion of the FP is depicted as a function of degree for synthetic networks.

Now let us focus on the solid black curve in Figure 2a which pertains to mean follower superiority. This proportion also increases initially, followed by a plateau that lasts up to degrees around 100, and then decreases. The decline is steeper than the case of mean followee superiority.

For the next example, we consider Figure 2h. The solid red curve starts at 0.3. It increases up to 0.9, with a plateau that lasts up to an FTR of around 0.04, then decreases monotonically. For example, a user with an FTR of 0.01 has a higher proportion of experiencing mean followee superiority in FTR (proportion is around 0.9) than a user with an FTR of 0.001 (for whom the proportion is 0.7).

The presence of positive slopes and/or plateaus is visible in most of the curves presented in Figure 2. Such a behavior is in stark contrast with one might intuitively expect to observe (a monotonically decreasing curve, as mentioned above). In Figures 2a, 2b, 2g, and 2h, the initial increase in the proportion of experiencing neighbor superiority indicates that those with minimum (and close to minimum) values of attributes establish links both amongst themselves and those with higher values of attributes. However, those with intermediate levels of the attributes tend to establish links only towards those who have higher attributes than them. In the hierarchical representation of the system, we can say that nodes with

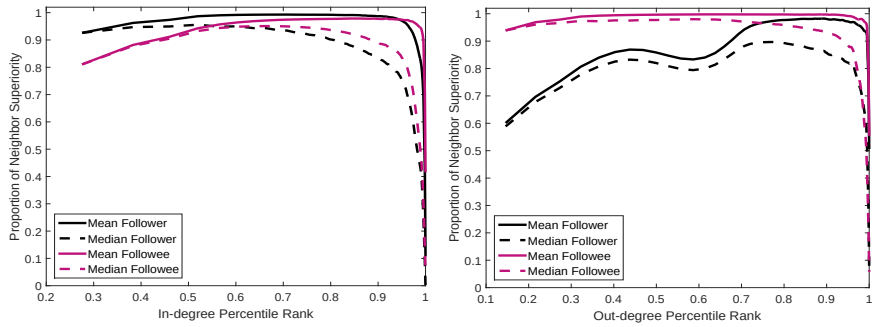
close-to-minimum attributes connect both up and across, and nodes with higher levels of attributes only connect up. We will get back to this point below when we discuss Figure 3.

In Figures 2c,2d, 2e, and 2f, the curves do not exhibit a steep positive slope. Rather, they begin with plateaus, followed by steep decrease. This pattern suggests that nodes with close-to-minimum levels of these attributes follow those with higher attributes than them. The hierarchies that pertain to these attributes are more upwardly-oriented; most nodes tend to connect up, rather than across.

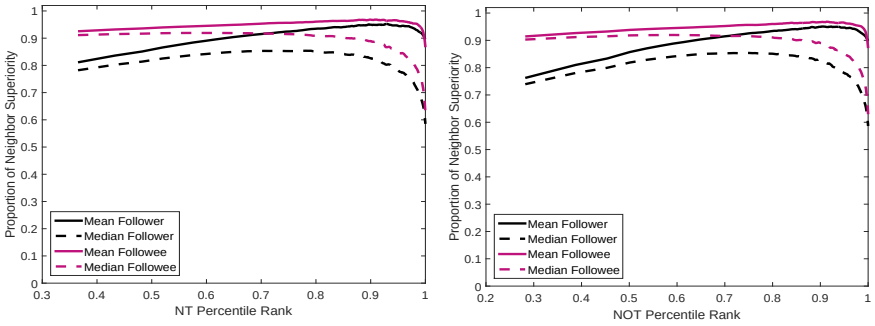
In all the figures, for the same type of superiority, the median curve falls below the mean curve. This implies that for any type of superiority, the proportion of experiencing median superiority is smaller than the proportion of experiencing mean superiority.

In Figure 2, different numbers of nodes fall into different bins. This is true for all nodal attributes. This is caused by the high skew in the distributions of the nodal attributes. This results in loss of valuable information. For example, we know from Figure 2e that a user whose TTR equals 10 experiences mean followee superiority with proportion of almost 0.9, that is, almost 90% of the users whose TTR is 10 experience mean followee superiority. However, this figure does not tell us where such users fall in the ranking of the TTR values.

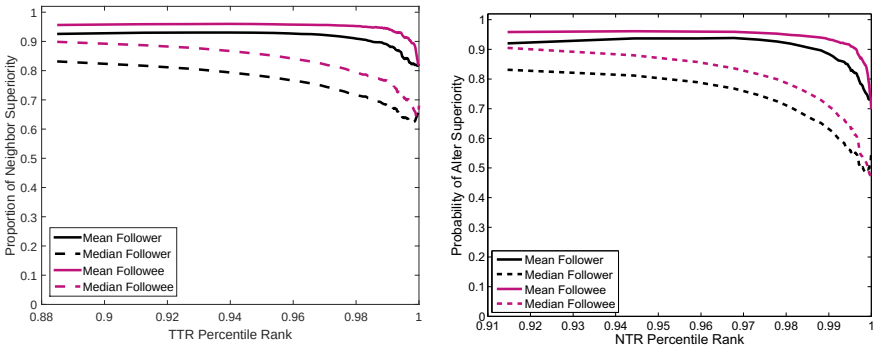
In Figure 3, we plot the proportion of experiencing different types of neighbor superiority as a function of the nodes' percentile rank for different attributes. The horizontal axes represent the percentile ranks. For example, in the case of NOT which is depicted in Figure 3d, a percentile rank of 0.6 for a node means that 60% of the nodes have NOT values smaller than or equal to the NOT of that node. In other words, the horizontal axes of Figure 3 results from a nonlinear rescaling of the horizontal axes of Figure 2. We have divided the interval between the minimum value and the maximum values for percentile ranks into 500 bins, and for nodes falling into the same bin, we calculated the fraction who experience the types of neighbor superiority corresponding to that attribute. The curves in Figure 3 are more telling: we readily observe that in all the figures, there is a very wide plateau that stretches up to the very close proximity of percentile rank of 1. This strengthens the assertions made above about the hierarchical nature of connections: most users—even those with very high ranking of any attribute—are connected to those with



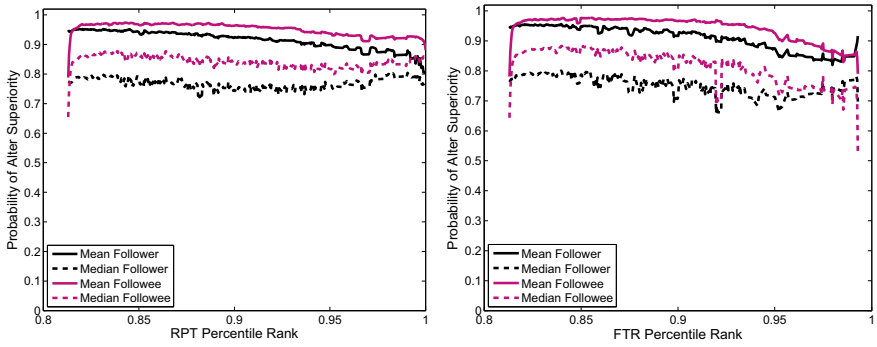
(a) In-degree **(b)** Out-degree



(c) NT: Number of Tweets **(d)** NOT: Number of Original Tweets.



(e) TTR: Total Times Retweeted. **(f)** NTR: Number of Tweets Retweeted.



(g) RPT: Retweets Per Tweet. **(h)** FTR: Fraction of Tweets Retweeted.

Figure 3. Proportion of neighbor superiority of different types for percentiles of nodal attributes.

higher attributes than them. Note that in some figures, the last bin does not have zero proportion of experiencing neighbor superiority. This is because the last bin stores the top 0.2% of the population, and the proportion associated with this bin is averaged among the top 0.2%. Since many of these users experience neighbor superiority, despite their very high ranking in the corresponding attributes, the average is not zero.

Figure 4 reaffirms the hierarchical nature of the connections. The figure is depicted as follows. We first divide the range of in-degrees into 25 logarithmic bins and group the population accordingly. We then construct matrix A whose (i,j) element denotes the number of links that are from a node in bin j to a node in bin i. We then normalize the matrix A column-wise, so that each column sums up to unity. Let us denote the resulting matrix by B. Matrix B is depicted in Figure 4. Column c represents the distribution of the destination of links whose starting points are nodes in bin c. The values on the bottom and left axes are the starting points of the bins. The values on the top and right axes are the corresponding percentile ranks of the starting points of the bins.

The matrix can be divided into four regions. For the nodes with in-degrees in the first five bins, the majority of outgoing links land on the nodes with highest in-degrees. For the nodes in the next nine bins the majority of the outgoing links point towards the nodes within the same bins. However, the fraction of links pointing to the nodes with larger in-degrees is higher than those pointing to nodes with smaller or equal in-degrees. The nodes in these two regions are likely to experience both the mean and median followee superiority. The nodes in the third region are mostly following the nodes with relatively high in-degrees (in the 98th and 99th percentiles), but not the highest bins. Finally, the nodes in the last four bins tend to follow nodes in the first seven bins.

Note that nodes in the second region experience the followee superiority because of following nodes within the same region but with higher in-degrees; not for following the hubs. Moreover, if the structure of the graph was star-like, we would expect to see two dense regions in the top left and bottom right of the matrix. This would disregard the role of the nodes in the middle bins.

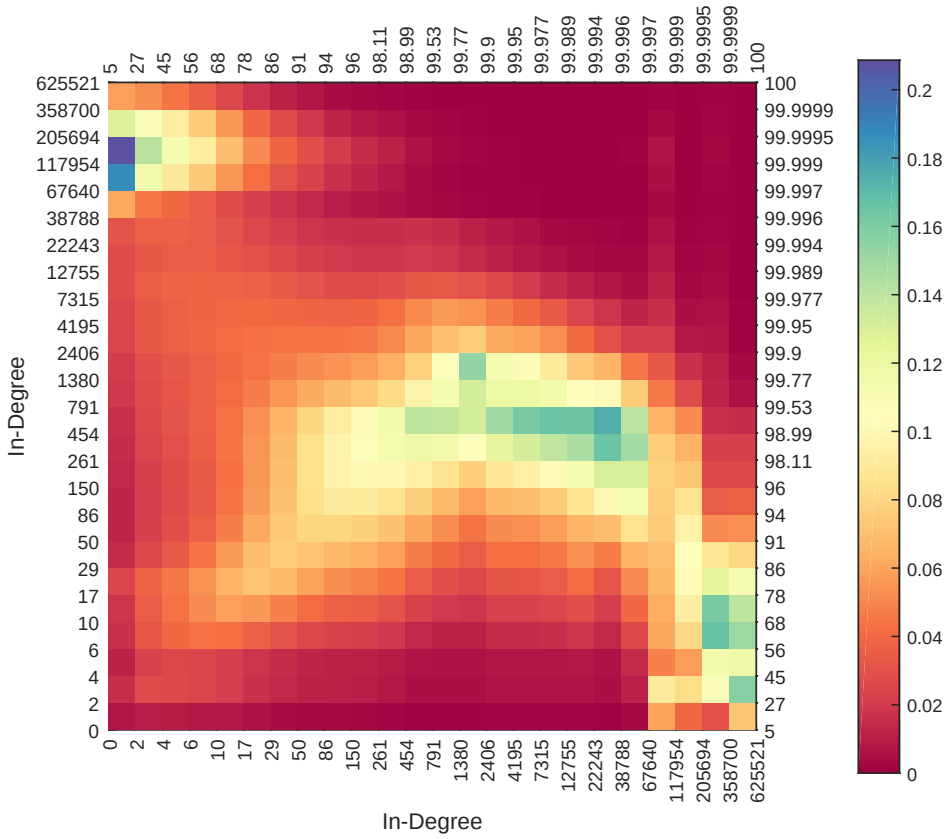


Figure 4. In-degree distribution of the followees of nodes as a function of their in-degree. The range of in-degrees are divided into 25 logarithmic bins. The values on the bottom and left axes are the starting points of the bins. The values on the top and right axes are the corresponding percentile ranks of the starting points of the bins. Each column is normalized.

Conclusion

In this paper we introduced six new measures to quantify different aspects of user activity and influence on social networks, and we computed them on a dataset of over 200 million tweets. We demonstrated that the distributions of all of these attributes are heavy-tailed. Two of these attributes (NT and NOT) are measures of activity, and four of them (TTR, NTR, RPT and FTR) pertain to received retweets, and are measures of influence. The measures of influence are zero for more than 75% of users, suggesting that the majority

of Twitter users are observers of the content produced by a minority.

We also introduced measures of neighbor superiority to quantify the local inequalities in different nodal attributes. We observed that the prevalence of mean neighbor superiorities of all types are above 63%. The prevalence of median neighbor superiorities of different types are also high; in 12 out of 16 types of median neighbor superiority, the prevalence is over 57%. We discussed that the high prevalence of median versions of neighbor superiority challenges the simplistic picture that neighbor superiority is a mere consequence of the existence of a few hubs in the network that put every peripheral node into experiencing neighbor superiority by elevating the average.

By inspecting different types of neighbor superiority, we uncovered the hierarchical nature of the connections in the Twitter graph both in terms of connectivity and in terms of nodal qualities. We observed that the fraction of nodes experiencing followee superiority exceeds the fraction of nodes experiencing follower superiority, and this is true for 15 out of 16 types of superiority introduced. This indicates the tendency of most users to follow other users who have higher attributes. It is of note that when we speak of hierarchical structures, there are distinct hierarchies for different attributes. That is, for example, if we once sort the node in terms of TTR, and then sort them in terms of in-degree, the hierarchies differ. Because the intra-node correlation between attributes are small (as presented in Table II) and therefore, a node that stands on the top of the hierarchy for TTR might be elsewhere for in-degree. Let us point out that there are two distinct patterns of correlations: inter-node correlations for a given attribute, and intra-node correlation of different attributes. Our results indicate that hierarchies stem from high inter-node correlations of each given attribute. In [9], it is shown that intra-node correlation between degree and attributes is *sufficient* for observing the GFP, and our result is that it is not *necessary*.

By close inspection of measures of neighbor superiority and the dependence of the likelihood of experiencing neighbor superiority on different attributes, we deduced that most users rarely follow down, rather, they tend to follow up or across, that is, they tend to follow other users with similar or higher attributes. This is true for almost every user, which makes even those in the top 0.5% of the population experience neighbor superiority

of different types. This means that Twitter does not possess a simple star-like structure, but is decentralized and inequalities exist locally for almost all nodes.

A counter-intuitive finding is that the proportion of experiencing neighbor superiority is not a monotonically-decreasing function of nodal attributes, or their ranks in those attributes. For example, it is not the case that the more re-tweets one receives, one's likelihood of experiencing neighbor superiority decreases. Rather, this likelihood is roughly constant up to the top percentile of the population in terms of retweets received. The trend is even reversed in the case of in-degree. For example, the proportion of experiencing neighbor superiority can even increase as in-degree increases. To ensure low likelihood of experiencing neighbor superiority, it does not suffice to increase one's attribute; one needs to stand in a very high percentile of the population.

I. Acknowledgments

This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada grant RGPAS 429296-2012.

References

[1] Feld SL. Why your friends have more friends than you do. American Journal of Sociology. 1991;p. 1464–1477.

[2] Ugander J, Karrer B, Backstrom L, Marlow C. The anatomy of the facebook social graph. arXiv preprint arXiv:11114503. 2011;.

[3] Hodas NO, Kooti F, Lerman K. Friendship paradox redux: Your friends are more interesting than you. arXiv preprint arXiv:13043480. 2013;.

[4] Christakis NA, Fowler JH. Social network sensors for early detection of contagious outbreaks. PloS one. 2010;5(9):e12948.

[5] Garcia-Herranz M, Moro E, Cebrian M, Christakis NA, Fowler JH. Using friends as sensors to detect global-scale contagious outbreaks. PloS one. 2014;9(4):e92413.

[6] Eom YH, Jo HH. Generalized friendship paradox in complex networks: The case of scientific collaboration. Scientific reports. 2014;4.

[7] Han B, Srinivasan A. Your friends have more friends than you do: identifying influential mobile users through random walks. In: Proceedings of the thirteenth ACM international symposium on Mobile Ad Hoc Networking and Computing. ACM; 2012. p. 5–14.

[8] Lattanzi S, Singer Y. The power of random neighbors in social networks. WSDM; 2015.

[9] Jo HH, Eom YH. Generalized friendship paradox in networks with tunable degree-attribute correlation. Physical Review E. 2014;90(2):022809.

[10] Fotouhi B, Momeni N, Rabbat MG. Generalized Friendship Paradox: An Analytical Approach. In: Social Informatics. Springer; 2014. p. 339–352.

[11] Kimura M, Saito K, Nakano R. Extracting influential nodes for information diffusion on a social network. In: AAAI. vol. 7; 2007. p. 1371–1376.

[12] Kim DA, Hwong AR, Stafford D, Hughes DA, O'Malley AJ, Fowler JH, et al. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. The Lancet. 2015;.

[13] Kimura M, Saito K, Nakano R, Motoda H. Extracting influential nodes on a social network for information diffusion. Data Mining and Knowledge Discovery. 2010;20(1):70–97.

[14] Budak C, Agrawal D, El Abbadi A. Limiting the spread of misinformation in social networks. In: Proceedings of the 20th international conference on World wide web. ACM; 2011. p. 665–674.

[15] Nguyen NP, Yan G, Thai MT, Eidenbenz S. Containment of misinformation spread in online social networks. In: Proceedings of the 4th Annual ACM Web Science Conference. ACM; 2012. p. 213–222.

[16] Droz M, Szwabiński J, Szabó G. Motion of influential players can support cooperation in prisoners dilemma. The European Physical Journal B-Condensed Matter and Complex Systems. 2009;71(4):579–585.

[17] Szolnoki A, Perc M. Resolving social dilemmas on evolving random networks. EPL (Europhysics Letters). 2009;86(3):30007.

[18] Hill S, Provost F, Volinsky C. Network-based marketing: Identifying likely adopters via consumer networks. Statistical Science. 2006;p. 256–276.

[19] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2003. p. 137–146.

[20] Perisic A, Bauch CT. Social contact networks and disease eradicability under voluntary vaccination. PLoS computational biology. 2009;5(2):e1000280.

[21] Song X, Chi Y, Hino K, Tseng B. Identifying opinion leaders in the blogosphere. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM; 2007. p. 971–974.

[22] Hodas NO. How limited visibility and divided attention constrain social contagion. In: In SocialCom. Citeseer; 2012. .

[23] Fotouhi B. Complex Networks: Dynamism of Connectivity and Opinion. McGill University, Montreal; 2014.

[24] Chou HTG, Edge N. They are happier and having better lives than I am: the impact of using Facebook on perceptions of others' lives. Cyberpsychology, Behavior, and Social Networking. 2012;15(2):117–121.

[25] Yang J, Leskovec J. Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on Web search and data mining. ACM; 2011. p. 177–186.

[26] Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? In: WWW '10: Proceedings of the 19th international conference on World wide web. New York, NY, USA: ACM; 2010. p. 591–600.

[27] Momeni N, Rabbat MG. Measuring the Generalized Friendship Paradox in Networks with Quality-Dependent Connectivity. In: Complex Networks VI. Springer; 2015. p. 45–55.

[28] Kooti F, Hodas NO, Lerman K. Network Weirdness: Exploring the Origins of Network Paradoxes. arXiv preprint arXiv:14037242. 2014;.

CHAPTER 9

Paper: Generalized Friendship Paradox in Growing Networks

The material presented in this chapter was published in the following proceedings:

B. Fotouhi, N. Momeni, M. Rabbat, "Generalized Friendship Paradox: An Analytical Approach", International Conference on Social Informatics. Springer International Publishing, 2014

The candidate's contribution in this paper are: contributing to the study design, devising all the proposed measures, conducting all the numerical experiments and simulations, and contributing to the writing of the main text.

Please note that the references of the manuscript are listed at the end of this chapter.

Generalized Friendship Paradox: An Analytical Approach

Babak Fotouhi, Naghmeh Momeni, and Michael G. Rabbat

Abstract

The friendship paradox refers to the sociological observation that, while the people's assessment of their own popularity is typically self-aggrandizing, in reality they are less popular than their friends. The generalized friendship paradox is the average alter superiority observed empirically in social settings, scientific collaboration networks, as well as online social media. We posit a quality-based network growth model in which the chance for a node to receive new links depends both on its degree and a quality parameter. Nodes are assigned qualities the first time they join the network, and these do not change over time. We analyse the model theoretically, finding expressions for the joint degree-quality distribution and nearest-neighbor distribution. We then demonstrate that this model exhibits both the friendship paradox and the generalized friendship paradox at the network level, regardless of the distribution of qualities. We also show that, in the proposed model, the degree and quality of each node are positively correlated regardless of how node qualities are distributed.

I. Introduction

The friendship paradox is a phenomenon observed in various social networks. The term was coined by Feld [1]. It has been empirically observed that people's perception of their own popularity is self-aggrandizing; most people believe that they are more popular than their friends on average [2]. However, Feld observed that in reality, most people have fewer friends than their friends do. In [3], this phenomena is used for the early detection of flu outbreaks among college students. In [4], it is utilized to efficiently sample early-warning sensors during catastrophic events such as hurricanes.

In addition to degree, the same paradox has been observed about other individual attributes (called the *generalized friendship paradox* [5], or GFP). For example, in [6] it has been observed that on Twitter, for most people, their friends share, on average, more viral content and also tweet more. In [5], it has been observed that in scientific collaboration networks, one's co-authors have, on average, more citations, more publications and more co-authors.

In this paper, we consider a network growth model which is a generalization of the preferential attachment scheme [7]. In our model, nodes are endowed with 'qualities' (ak.a. 'fitness' or 'attractiveness' in the literature [8]–[11]). Qualities are discrete positive numbers drawn from a given distribution $\rho(\theta)$ and assigned to a node upon its birth (remaining the same thenafter). We assume that the probability that node x with degree k_x and quality θ_x receives a link from subsequent nodes is proportional to $k_x + \theta_x$.[1] We obtain two statistical measures of this model: one is the degree-quality joint distribution, which is the fraction of nodes that have degree k and quality θ in the steady state. The second quantity is the nearest-neighbor distribution of quality and degree: it gives the fraction of nodes with degree ℓ and quality ϕ that are connected to a node with degree k and quality θ. Equipped with these distributions, we can quantify the paradox and study how it depends on the underlying quality distribution $\rho(\theta)$. To our knowledge, no similar theoretical result is available in the literature for any network growth model (either purely preferential [7], or fitness-based [9]–[11]).

We show that employing the above scheme as the attachment mechanism renders the occurrence of the GFP contingent upon the underlying distribution of node qualities. We then employ measures defined in the literature for assessing the GFP on the network level, and we investigate the dependence of these measures on the model parameters and the quality distribution. We demonstrate that, in the proposed model, the network exhibits a quality paradox at the network level for any quality distribution. We contend that this

[1]Note that for example in [8], the attachment probability is proportional to the product of degree and quality. This model however, has not be solved in closed form. Also, it assigns zero link reception probability to nodes with degree zero.

is indicative of a positive correlation between degree and quality; i.e., those with higher qualities are more likely to have higher degrees, and vice versa.

II. Model, Notation and Terminology

In the growth model considered in this paper, nodes are added successively to the network. The initial network has $N(0)$ nodes and $L(0)$ links. At each time step, one new node is added to the network. We assume that each node has an intrinsic quality, which is drawn from a given distribution $\rho(\theta)$. The quality is assigned to each new incoming node upon birth, and will remain the same thenafter. The mean of the distribution $\rho(\theta)$ is denoted by μ. A node of degree k and quality θ is also referred to as *a* (k, θ) *node* throughout.

Each new incoming node attaches to $\beta \leq N(0)$ existing nodes in the network. We consider the simplest additive model that incorporates both degree (popularity) and quality in the dynamics of connection formation: the probability that an existing node with degree k and quality θ receives a link from the new node is proportional to $k + \theta$. This means that, for example, a paper that is new and has very few citations can compensate for its small degree with having a high quality. Or in the social context, a newcomer who does not have many friends in the new social milieu but is gregarious and sociable can elevate the chances of making new friends. The new node is called the *child* of the existing nodes that it connects to, and they are called its *parents*. By *a* (ℓ, ϕ)-(k, θ) *child-parent pair*, we mean a node with degree ℓ and quality ϕ that is connected to a parent node of degree k and quality θ.

The probability that an existing node x receives a new link is $\frac{k_x+\theta_x}{A}$, where the normalization factor A is given by $\sum_x (k_x + \theta_x)$. The sum over all node degrees at time t, which equals twice the number of links at time t, is equal to $2[L(0) + \beta t]$. For long times, the sum over the quality values of all the nodes will converge to the mean of the quality distribution times the number of nodes, that is, we can replace $\sum_x \theta_x$ by $[N(0) + t]\mu$. So at time t, the probability that node x receives a link equals $\frac{k_x+\theta_x}{2L(0)+N(0)+(2\beta+\mu)t}$.

Throughout the present paper, the steady-state joint distribution of quality and degree is denoted by $P(k, \theta)$. The expected number of nodes with degree k and quality θ at time

t is denoted by $N_t(k,\theta)$. We denote by $N_t(k,\theta,\ell,\phi)$ the expected number of (ℓ,ϕ)-(k,θ) child-parent pairs.

III. Degree-quality Joint Distribution

We seek the steady-state fraction of nodes who have degree k and quality θ. In Appendix A we derive the following expression for this quantity:

$$P(k,\theta) = \rho(\theta)\left(2+\frac{\mu}{\beta}\right)\frac{\Gamma(k+\theta)}{\Gamma(\beta+\theta)}\frac{\Gamma\left(\beta+\theta+2+\frac{\mu}{\beta}\right)}{\Gamma\left(k+\theta+3+\frac{\mu}{\beta}\right)}u(k-\beta). \tag{1}$$

Note that in the special case of a single permitted value for the quality (that is, when $\rho(\theta) = \delta[\theta-\theta_0]$) this model reduces to the shifted-linear preferential attachment model analyzed, for example, in [12]. The solution in this special case simplifies to

$$P_{sh}(k) = \left(2+\frac{\theta_0}{\beta}\right)\frac{\Gamma(k+\theta_0)}{\Gamma(\beta+\theta_0)}\frac{\Gamma(\beta+2+\theta_0+\frac{\theta_0}{\beta})}{\Gamma(k+3+\theta_0+\frac{\theta_0}{\beta})}. \tag{2}$$

This coincides with the degree distribution of shifted-linear kernels given in [13] and [12, Equation D.9]. Furthermore, when $\rho(0) = 1$, all nodes will have zero quality and attachments will be purely degree-proportional, synonymous with the conventional preferential-attachment model proposed initially in [7]. For the special case of $\theta = \mu = 0$ we obtain

$$P_{BA}(k) = \frac{2\beta(\beta+1)}{k(k+1)(k+2)}. \tag{3}$$

This is equal to the degree distribution of the conventional BA network (see, e.g., [13], [14]).

Let us also examine the behavior of (1) in the limit of large k. In this regime, we can use the asymptotic approximation that for large values of x, the function $\Gamma(x) \approx x^{x-\frac{1}{2}}\exp(-x)$. Then we replace $\frac{\Gamma(k+\theta)}{\Gamma(k+\theta+3+\frac{\mu}{\beta})}$ with $k^{-3-\frac{\mu}{\beta}}$, independent of θ. Therefore, the steady-state joint degree-quality distribution $P(k,\theta)$ is proportional to $k^{-3-\frac{\mu}{\beta}}$. Marginalizing out θ to recover the degree distribution, we obtain the well-known power law, $P(k) = k^{-3-\frac{\mu}{\beta}}$.

IV. Nearest-Neighbor Quality-Degree Distribution

To quantify how qualities and degrees of adjacent nodes correlate, we need to go beyond the quality-degree distribution obtained in the previous section. The closed-form expression for the nearest-neigbor correlations under the preferential attachment model is derived in [12]; that work only considers degrees and does not address qualities. We would like to quantify the conditional distribution $P(\ell, \phi|k, \theta)$, the fraction of neighbours of a given node with degree k and quality θ that have degree ℓ and quality ϕ. We refer to this as the *nearest-neighbor quality-degree distribution* (NNQDD).

In Appendix A we study the rate equation describing how the distribution $P(\ell, \phi|k, \theta)$ evolves as nodes are added to the network. This gives rise to a system of difference equations which we solve to obtain that, in the steady-state,

$$P(\ell,\phi|k,\theta) = \frac{\rho(\phi)}{k}\frac{\Gamma\left(k+\theta+3+\frac{\mu}{\beta}\right)}{\Gamma\left(k+\theta+3+\frac{\mu}{\beta}+\ell+\phi\right)}\frac{(\ell-1+\phi)!}{(\beta-1+\phi)!}\Gamma\left(\beta+2+\phi+\frac{\mu}{\beta}\right)\times$$
$$\left[\sum_{j=\beta+1}^{k}\frac{\Gamma\left(j+\theta+2+\frac{\mu}{\beta}+\beta+\phi\right)\binom{k-j+\ell-\beta}{\ell-\beta}}{\Gamma\left(j+\theta+2+\frac{\mu}{\beta}\right)\Gamma\left(\beta+2+\phi+\frac{\mu}{\beta}\right)}+\sum_{j=\beta+1}^{\ell}\frac{\Gamma\left(j+\theta+2+\frac{\mu}{\beta}+\beta+\phi\right)\binom{\ell-j+k-\beta}{k-\beta}}{\Gamma\left(j+\phi+2+\frac{\mu}{\beta}\right)\Gamma\left(\beta+2+\theta+\frac{\mu}{\beta}\right)}\right]. \tag{4}$$

In order to obtain the nearest-neighbor quality distribution $P(\phi|\theta)$, one needs to perform the calculations $P(\phi|\theta) = \sum_\ell \sum_k P(k)P(\ell, \phi|k, \theta)$, which requires knowledge of $P(k)$. In turn we have $P(k) = \sum_\theta P(k, \theta)$, which according to (1), yields different sums for different quality distributions $\rho(\theta)$.

V. Quantifying the Friendship and Generalized Friendship Paradoxes

As discussed in Section I, GFP refers to an average alter superiority in arbitrary aspects (e.g., number of citations, exposure to viral online content). In this paper, we use the 'quality' dimension that is incorporated in the model as the subject of the GFP. Our objective is to compare the degrees and qualities of nodes with their neighbors. We say that a node experiences the friendship paradox if the degree of that node is less than the average of

the degrees of its neighbors. Similarly, we say that a node experiences the quality paradox if the quality of the node is less than the average of the qualities of its neighbors.

The above-mentioned definitions characterize individual-level paradoxes. Our primary interest is to what fraction of nodes experience the friendship and quality paradoxes. To this end, we compare the average degree of the nodes with the average degree of the neighbors of all nodes (and similarly for quality). Comparing these two average values yields a macro measure for the system, indicating whether it exhibits paradoxes on average. We call these as the *network-level friendship paradox* and *network-level quality paradox*.

Our measure of the network-level quality paradox is defined as NQP $= \frac{\sum_i k_i \theta_i}{\sum_i k_i} - \frac{1}{N} \sum_i \theta_i$. The summations are performed over all nodes in the network. Note that the numerator of the first sum is actually the sum of the qualities of the neighbors of all nodes. Node i is repeated k_i times in this sum, once for each of its neighbors. Focusing on the limit as $t \to \infty$, we can use the law of large numbers and express the NQP as follows

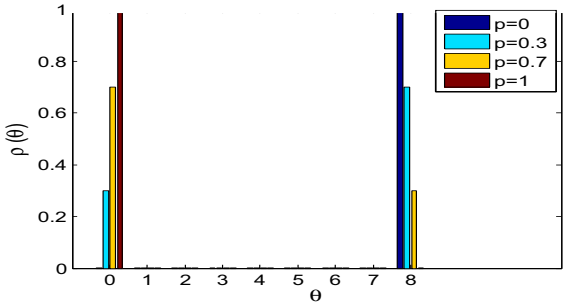
$$\text{NQP} = \frac{\sum_{k,\theta} k\theta P(k,\theta)}{\sum_{k,\theta} kP(k,\theta)} - \mu. \tag{5}$$

The greater NQP becomes, the more strongly the paradox holds. Negative NQP is indicative of the absence of a quality paradox at the network level.

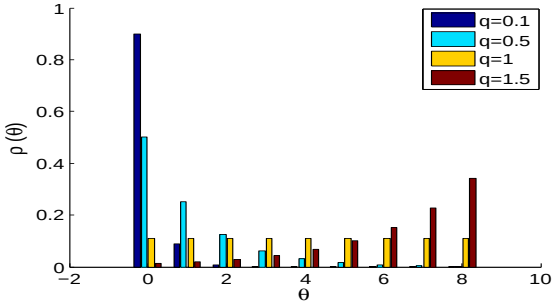
Undertaking similar steps to above, we can measure the network-level friendship paradox via

$$\text{NFP} = \frac{\langle k^2 \rangle}{\langle k \rangle} - \langle k \rangle = \frac{\langle k^2 \rangle - \langle k \rangle^2}{\langle k \rangle}. \tag{6}$$

Note that the numerator is the variance of the degree distribution, so it is positive. The denominator is the average degree and is also positive. So the NFP is always positive, which means that by this definition: *any network exhibits the friendship paradox at the network level.* So the task of the present paper with regard to the NFP is to investigate its magnitude, i.e., to measure how strongly the paradox holds. For example, in the conventional Barabasi-Albert scale-free model, where the degree variance diverges, the NFP also diverges, which is a result of the presence of macro hubs.



(a) Bernoulli distribution with $p = 0, 0.3, 0.7, 0.1$. The cases of $p = 0$ and $p = 1$ correspond to conventional Barabasi-Albert and shifted-linear preferential attachment networks, respectively.



(b) Exponential distribution for decay factor $q = 0.1, 0.5, 1, 1.5$. The special case of $q = 1$ corresponds to a uniform distribution supported in the interval $0 \leq \theta \leq \theta_{\max}$.

Fig. 1: Examples of the quality distributions used in this paper with $\theta_{\max} = 8$. Four instances of each type is depicted.

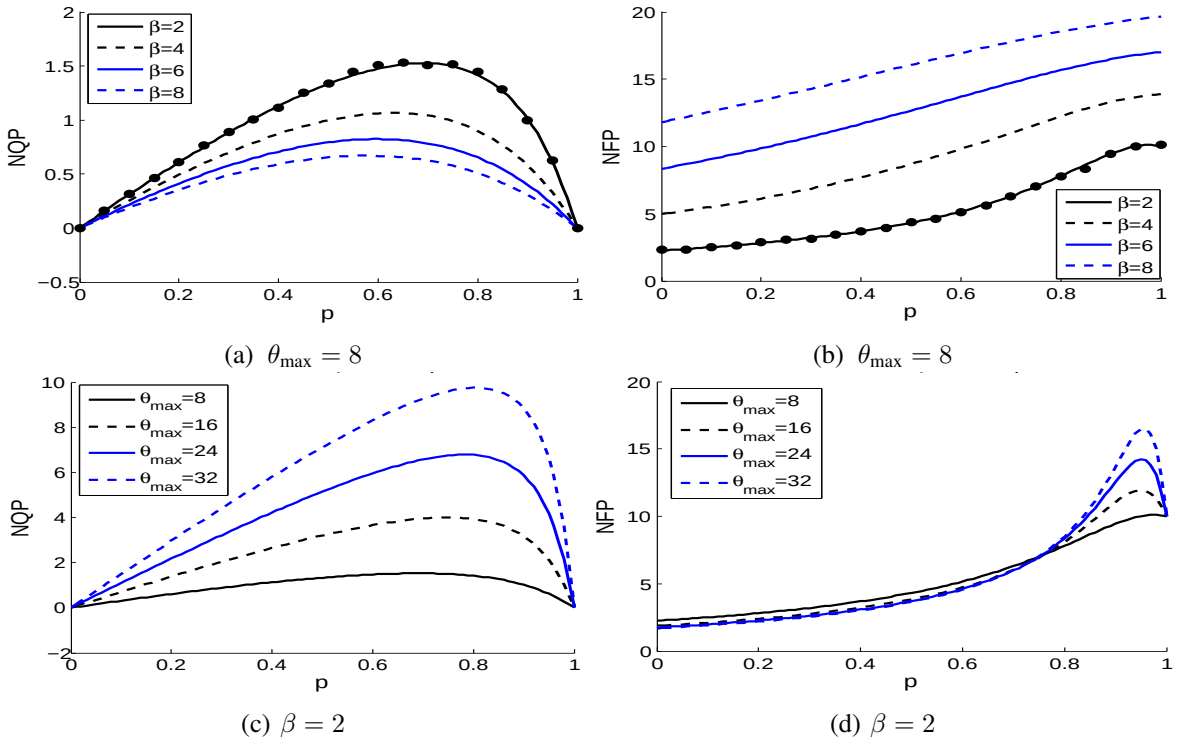
VI. Results and Discussion

To study the NFP and the NQP in concrete settings, we confine ourselves to two quality distributions $\rho(\theta)$ for illustrative purposes. We consider a finite support for θ, so that $0 \leq \theta \leq \theta_{\max}$. For each distribution, we are going to consider four different values β, and four different values of $\theta_{\max}$.

The first distribution we consider is the Bernoulli case, where nodes can either have quality zero or quality $\theta_{\max}$. The probability of quality zero is p and the probability of quality $\theta_{\max}$ is $1 - p$, where $0 \leq p \leq 1$. The second distribution we consider is the discrete exponential distribution with decay factor q. The probability that the quality is θ is proportional to q^{θ}. Note that in the case of $q = 1$, one recovers a uniform distribution as a special case. We consider both $q < 1$ and $q > 1$, yielding decreasing and increasing distributions in θ, respectively. These distributions are depicted in Figure 1.

The results for the Bernoulli quality distribution are depicted in Figure 2. As depicted in Figure 2a, for a fixed $\theta_{\max}$, the NQP decreases as β (the initial degree of nodes) increases. Also, it is observable that the sensitivity of the NQP to the variations of the quality distribution diminishes for larger values of β.

As illustrated in Figure 2b, the NFP increases as β (the initial degree of nodes) increases.



(a) $\theta_{\max} = 8$ (b) $\theta_{\max} = 8$

(c) $\beta = 2$ (d) $\beta = 2$

Fig. 2: Network level friendship and quality paradox for Bernoulli quality distribution. The markers in Figures (a) and (b) represent simulation results, and the solid curves are the theoretical expression. The depicted results are averaged over 100 Monte Carlo trials.

Hence, according to (6) *the variance of the degree distribution grows faster than the mean degree, as* β *increases.* On the other hand, for a given β, increasing $\theta_{\max}$ (which is tantamount to increasing μ), increases the NQP. This means that according to (5) *as* $\theta_{\max}$ *increases, the mean of the qualities of the neighbors increases faster than the mean of the qualities of the nodes.*

Figure 2c pertains to this case. Observe that as $\theta_{\max}$ increases, the NQP becomes more sensitive to the distribution of qualities. Finally, Figure 2d represents the NFP for a fixed β and different values of $\theta_{\max}$. From Figures 2a, 2b, 2c and 2d, a general observable pattern is that as p increases, the NFP increases (monotonically for almost all values of p), whereas the NQP is concave and unimodal (it increases at first, achieves maximum, and then decreases).

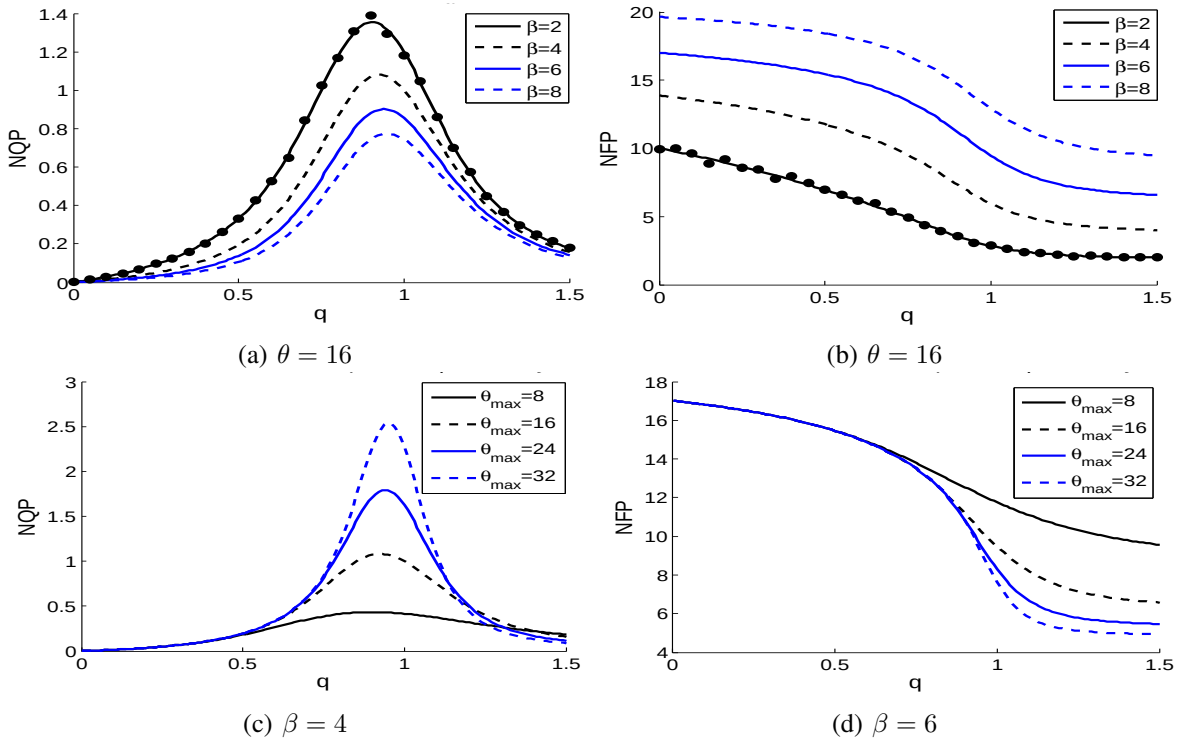
Now we focus on the exponential quality distribution with the decay factor denoted by q. As depicted in Figure 3a, for a given θ_{max}, the NQP decreases as β increases. Also, it is observed that as β increases, the sensitivity of the NQP to the quality distribution diminishes. These are both similar to the results of the Bernoulli distribution. As can be seen in Figure 3b, the NFP increases as β increases. So similar to the Bernoulli case, the variance of the degree distribution grows faster than the mean degree, as β increases.

From Figure 3c we observe that for a fixed β, increasing θ_{max} increases the NQP. We observe that as θ_{max} increases, NQP becomes more sensitive to the changes in the decay factor. Finally, Figure 3d represents the NFP for a fixed β and different values of θ_{max}. We observe that increasing θ_{max} increases the NFP for positive decays. Also, for very small decay factors (which generate right-skewed distributions that are highly unequal), changing θ_{max} has scant effect on the NFP. This is reasonable because when the decay factor is small, all large values of θ have small chances of occurrence. Consequently, changing θ_{max} minimally changes the shape of the distribution for small decay factors.

A trend is discernible from Figures 3a, 3b, 3c and 3d: as q increases, the NFP decreases (monotonically for all values of q), whereas NQP is concave and increases up to a point around $q = 1$, and then decreases. Since $q = 1$ yields a uniform distribution, we can qualitatively conclude that the probability of the network-level quality paradox is higher when qualities are heterogeneous, as compared to when qualities are similar.

Finally, to verify our results, we run Monte Carlo simulations to synthesize networks that grow under the prescribed quality-based preferential attachment mechanism, and then calculate the desired quantities by averaging over nodes in the synthesized network. Due to computational limitations, we restrict this validation to the case where $\beta = 2$ and $\theta_{\text{max}} = 8$ for the Bernoulli quality distribution and the case where $\beta = 2$ and $\theta_{\text{max}} = 16$ for the exponential quality distribution. These results are shown in Figures 2a , 2b, 3a and 3b. The markers show the results of simulations, averaging over 100 Monte Carlo trials, and the solid curves correspond to our theoretical expressions.

We have tested the results on various other quality distributions and observed similar results; these additional simulations not reported here due to space limitations. In general,



(a) $\theta = 16$ (b) $\theta = 16$

(c) $\beta = 4$ (d) $\beta = 6$

Fig. 3: Network level friendship and quality paradox for exponential quality distribution. The markers in Figures (a) and (b) represent simulation results and the solid curves are from the theoretical expressions. The depicted results are averaged over 100 Monte Carlo trials.

we observe that for a fixed $\theta_{\max}$, increasing β increases the NFP and decreases the NQP regardless of the quality distribution. Also, for a fixed β, increasing $\theta_{\max}$ increases the NQP and decreases the NFP.

Note that in all cases the NQP is nonnegative. This has roots in the correlation between degree and quality of single nodes (intra-node correlation, rather than inter-node correlation). Let us denote the correlation between degree and quality for a node by $\rho_{k\theta}$, which is the Pearson correlation coefficient obtained from the joint distribution $P(k, \theta)$. From (5),

we have:

$$\begin{aligned}\text{NQP} &= \frac{\sum_{k,\theta} k\theta P(k,\theta)}{\sum_{k,\theta} kP(k,\theta)} - \mu = \frac{\sum_{k,\theta} k\theta P(k,\theta) - \mu \sum_{k,\theta} kP(k,\theta)}{\sum_{k,\theta} kP(k,\theta)} \\ &= \frac{\sum_{k,\theta} k\theta P(k,\theta) - \mu \langle k \rangle}{\langle k \rangle} = \frac{\rho_{k\theta}\sigma_k\sigma_\theta}{\langle k \rangle}. \end{aligned} \tag{7}$$

This implies that the sign of NQP is the same as the sign of $\rho_{k\theta}$ (since σ_k, σ_θ and $\langle k \rangle$ are nonnegative). The observation that NQP is always nonegative indicates that $\rho_{k\theta}$ is also always nonegative. We conclude that *the quality-dependent preferential attachment model generates networks in which degree and quality of a node are always positively correlated.* This is what we intuitively expect the model to exhibit; increasing quality increases degree. For example, in citation networks, papers with higher qualities receive more citations. Conversely, a paper with many citations is more likely to have a high quality. In the case of friendship networks, a person that is more sociable ends up with more friends than an anti-social person, and conversely, a popular person is more likely to be friendly than an isolated person.

We also observe that in all cases, μ (equivalently, θ_{max}) and β have opposite effects on both the NFP and the NQP. That is, the effect of increasing β is akin to that of decreasing μ, and vice versa. We observed similar trends for other quality distributions; these results are omitted here due to space limitations. What causes this disparity is the following: as can be seen in (1) and (22), μ only appears in the distributions in the form of $\frac{\mu}{\beta}$. Thus increasing μ and decreasing β have the same effect on this variable, and consequently, on the distribution.

VII. Summary and Future Work

The aim of the present paper was to put in crisp theoretical focus the seemingly prevalent phenomena of the friendship paradox and the generalized friendship paradox. We proposed a network growth model that incorporates quality. In this model, the probability that a node receives a link increases with both its degree and quality. We analysed the model theoretically in the steady-state (large size limit), and found two theoretical quantities that

characterize the interrelation between quality and degree. The first quantity is $P(k, \theta)$, which is the joint degree-quality distribution, and equals the fraction of nodes who have degree k and quality θ. The second quantity characterizes nearest-neighbor correlations, and is the nearest-neighbor quality-degree distribution, denoted by $P(\ell, \phi|k, \theta)$.

We then defined two network-level measures for the quality and friendship paradoxes and computed them for two particular examples of quality distributions. We observed that for a fixed θ_{max}, increasing β increases the NFP and decreases the NQP regardless of the quality distribution. We also observed that for a fixed β, increasing θ_{max} increases the NQP and decreases the NFP. We also observed that μ and β have opposite effects on the NFP and also on the NQP. We also tested these results on various other quality distributions, and they proved robust; the effects of β and μ on paradoxes are opposite regardless of the quality distribution.

There are many interesting extensions of this work to pursue. In addition to the network-level paradox, we can also study the individual-level paradox, which would require the utilization of the NNQDD to compare the degrees and qualities of nodes with those of their neighbors. The individual-level paradox has empirical implications which enable us to assess the quality distribution of real networks.

APPENDIX

We seek the fraction of nodes who have degree k and have quality θ. We begin by writing the rate equation which quantifies the temporal evolution of $N_t(k,\theta)$. Suppose that a node with quality θ and degree $k-1$ at time $t-1$, receives a link from the new incoming node. Consequently, its degree will become k and $N_t(k,\theta)$ increments. Conversely, if a node with quality θ and degree k at time $t-1$, receives a link from the new incoming node, $N_t(k,\theta)$ decrements. Finally, each new incoming node increments $N_t(\beta,\theta)$ with probability $\rho(\theta)$. The rate equation thus reads

$$N_{t+1}(k,\theta) - N_t(k,\theta) = \frac{\beta(k-1+\theta)N_t(k-1,\theta)}{2L(0)+N(0)+(2\beta+\mu)t} - \frac{\beta(k+\theta)N_t(k,\theta)}{2L(0)+N(0)+(2\beta+\mu)t} + \rho(\theta)\delta_{k,\beta}. \tag{8}$$

Replacing $N_t(k,\theta)$ by $[N(0)+t]P_t(k,\theta)$, this can be expressed in terms of $P_t(k,\theta)$ as follows:

$$\big[N(0)+t\big]\big[P_{t+1}(k,\theta)-P_t(k,\theta)\big]+P_{t+1}(k,\theta) = \frac{\beta(k-1+\theta)[N(0)+t]P_t(k-1,\theta)}{2L(0)+N(0)+(2\beta+\mu)t} - \frac{\beta(k+\theta)[N(0)+t]P_t(k,\theta)}{2L(0)+N(0)+(2\beta+\mu)t} + \rho(\theta)\delta_{k,\beta}. \tag{9}$$

In the limit as $t\to\infty$, the transients vanish. So, we drop the t in the arguments and rewrite (9) as:

$$P(k,\theta) = \frac{\beta(k-1+\theta)P(k-1,\theta)}{2\beta+\mu} - \frac{\beta(k+\theta)P(k,\theta)}{2\beta+\mu} + \rho(\theta)\delta_{k,\beta}. \tag{10}$$

This can be rearranged and expressed equivalently as follows:

$$P(k,\theta) = \frac{(k-1+\theta)P(k-1,\theta)}{2+\frac{\mu}{\beta}+k+\theta} + \frac{2+\frac{\mu}{\beta}}{2+\frac{\mu}{\beta}+\beta+\theta}\rho(\theta)\delta_{k,\beta}. \tag{11}$$

Multiplying both sides by $2\beta+\mu$ and rearranging the terms, this can be recast as follows

$$P(k,\theta) = \frac{(k-1+\theta)P(k-1,\theta)}{2+\frac{\mu}{\beta}+k+\theta} + \frac{2+\frac{\mu}{\beta}}{2+\frac{\mu}{\beta}+\beta+\theta}\rho(\theta)\delta_{k,\beta}. \tag{12}$$

Setting $k = \beta$, this yields $P(\beta, \theta) = \dfrac{2 + \frac{\mu}{\beta}}{2 + \frac{\mu}{\beta} + \beta + \theta}\ \rho(\theta)$. For all $k > \beta$, the second term on the right hand side vanishes, and this equation reduces to a straightforward recursion $P(k, \theta) = \dfrac{(k - 1 + \theta)}{2 + \frac{\mu}{\beta} + k + \theta} P(k - 1, \theta)$, whose solution is

$$P(k,\theta) = P(\beta,\theta) \prod_{j=\beta+1}^{k} \frac{(k-1+\theta)}{\left(2 + \dfrac{\mu}{\beta} + k + \theta\right)} = P(\beta,\theta)\, \frac{(k-1+\theta)!}{(\beta-1+\theta)!}\, \frac{\Gamma\left(3 + \dfrac{\mu}{\beta} + \beta + \theta\right)}{\Gamma\left(3 + \dfrac{\mu}{\beta} + \beta + \theta\right)}$$

$$= \rho(\theta) \left(2 + \frac{\mu}{\beta}\right) \frac{\Gamma(k+\theta)}{\Gamma(\beta+\theta)} \frac{\Gamma\left(\beta + 2 + \theta + \dfrac{\mu}{\beta}\right)}{\Gamma\left(k + 3 + \theta + \dfrac{\mu}{\beta}\right)}. \tag{13}$$

We begin by writing the rate equation to quantify the evolution of $N_t(k, \theta, \ell, \phi)$, which is the number of nodes with degree ℓ and quality ϕ who are connected to a parent node of degree k and quality θ. Upon introduction of a new node, regardless of its quality, the following is true: if it attaches to a node of degree ℓ and quality ϕ who is the child of a parent of degree k and quality θ, then the degree of the receiving node increments and consequently $N_t(k, \theta, \ell, \phi)$ decrements. Also, $N_t(k, \theta, \ell, \phi)$ decrements if the new node attaches to the parent node in such a pair of nodes. Another way that $N_t(k, \theta, \ell, \phi)$ can increment is if either there is a child-parent pair of $(k, \theta, \ell - 1, \phi)$ or $(k - 1, \theta, \ell, \phi)$. If the new node attaches to the child node in the former case or to the parent node in the latter case, then $N(k, \theta, \ell, \phi)$ increments. Finally, with probability $\rho(\phi)$, the new node will have quality ϕ, and if the new node attaches to an existing node of degree $k - 1$ and quality θ, then $N_t(k, \theta, \ell, \phi)$ increments. The rate equation reads

$$N_{t+1}(k,\theta,\ell,\phi) - N_t(k,\theta,\ell,\phi) = \beta \left[\frac{(\ell - 1 + \phi) N_t(k,\theta,\ell-1,\phi) - (\ell + \phi) N_t(k,\theta,\ell,\phi)}{2L(0) + N(0) + (2\beta + \mu)t} \right]$$
$$+\beta \left[\frac{(k - 1 + \theta) N_t(k-1,\theta,\ell,\phi) - (k + \theta) N_t(k,\theta,\ell,\phi)}{2L(0) + N(0) + (2\beta + \mu)t} \right] + \rho(\phi)\delta_{\ell,\beta}\, \frac{\beta(k - 1 + \theta) N_t(k-1,\theta)}{2L(0) + N(0) + (2\beta + \mu)t} \tag{14}$$

Undertaking the same steps that let us transform (8) into (9), and denoting the fraction $\frac{N(k,\theta,\ell,\phi)}{N(0)+t}$ by $n_t(k, \theta, \ell, \phi)$, this can be re-written in terms of $n_t(k, \theta, \ell, \phi)$ instead of

$N_t(k,\theta,\ell,\phi)$. In the limit as $t\to\infty$, we can drop the t subscript and obtain:

$$n(k,\theta,\ell,\phi) = \frac{(\ell-1+\phi)n(k,\theta,\ell-1,\phi)}{2+\frac{\mu}{\beta}+k+\ell+\theta+\phi} + \frac{(k-1+\theta)n(k-1,\theta,\ell,\phi)}{2+\frac{\mu}{\beta}+k+\ell+\theta+\phi} + \rho(\phi)\delta_{\ell,\beta}\frac{(k-1+\theta)P(k-1,\theta)}{2+\frac{\mu}{\beta}+k+\ell+\theta+\phi}. \tag{15}$$

Let us define the new sequence $m(k,\theta,\ell,\phi) = \frac{\Gamma(3+\frac{\mu}{\beta}+k+\ell+\theta+\phi)}{(k-1+\theta)!(\ell-1+\phi)!}n(k,\theta,\ell,\phi)$. Using this substitution and applying the properties of the Gamma function as well as the delta function, we can rewrite (15) equivalently as

$$m(k,\theta,\ell,\phi) = m(k,\theta,\ell-1,\phi) + m(k-1,\theta,\ell,\phi) + \frac{\Gamma\left(2+\frac{\mu}{\beta}+k+\beta+\theta+\phi\right)}{(k-1+\theta)!(\beta-1+\phi)!}\rho(\phi)\delta_{\ell,\beta}(k-1+\theta)P(k-1,\theta). \tag{16}$$

Using the expression in (1) to rewrite the last term on the right hand side of this equation, we can express it equivalently as follows

$$m(k,\theta,\ell,\phi) = m(k,\theta,\ell-1,\phi) + m(k-1,\theta,\ell,\phi) + \rho(\phi)\rho(\theta)\delta_{\ell,\beta}\left(2+\tfrac{\mu}{\beta}\right)\frac{\Gamma\left(2+\frac{\mu}{\beta}+k+\beta+\theta+\phi\right)}{(\beta-1+\theta)!(\beta-1+\phi)!}\frac{\Gamma\left(\beta+2+\theta+\frac{\mu}{\beta}\right)}{\Gamma\left(k+2+\theta+\frac{\mu}{\beta}\right)}. \tag{17}$$

Now define the generating function $\psi(z,\theta,y,\phi) = \sum_k m(k,\theta,\ell,\phi)z^{-k}y^{-\ell}$. Multiplying both sides of (16) by $z^{-k}y^{-\ell}$, summing over all values of k,ℓ and rearranging the terms, we arrive at

$$\psi(z,\theta,y,\phi) = \frac{\rho(\phi)\rho(\theta)\left(2+\frac{\mu}{\beta}\right)\Gamma\left(\beta+2+\theta+\frac{\mu}{\beta}\right)}{(\beta-1+\theta)!(\beta-1+\phi)!} \times \sum_{j=\beta+1}^{\infty}\frac{\Gamma\left(2+\frac{\mu}{\beta}+j+\beta+\theta+\phi\right)}{\Gamma\left(j+2+\theta+\frac{\mu}{\beta}\right)}\frac{z^{-j}y^{-\beta}}{1-z^{-1}-y^{-1}}. \tag{18}$$

(The lower bound of the sum is $\beta+1$ because $P(k-1,\theta)$ is zero for $k<\beta+1$.) The inverse transform of the factor $\frac{z^{-j}y^{-\beta}}{1-z^{-1}-y^{-1}}$in the summand can be taken through the following steps:

$$\begin{aligned}\frac{z^{-j}y^{-\beta}}{1-z^{-1}-y^{-1}} \xrightarrow{\mathcal{Z}^{-1}} & \frac{1}{(2\pi i)^2}\oint\oint \frac{z^{k-j-1}y^{\ell-\beta-1}}{1-z^{-1}-y^{-1}}dzdy \\ &= \frac{1}{(2\pi i)^2}\oint\oint \frac{z^{k-j}y^{\ell-\beta}}{z-\frac{y}{y-1}}\frac{1}{y-1}dzdy \\ &= \frac{1}{(2\pi i)}\oint\oint y^{\ell-\beta}\left(\frac{y}{y-1}\right)^{k-j}\frac{1}{y-1}dzdy \\ &= \frac{1}{(k-j)!}\left.\frac{d^{k-j}}{dy^{k-j}}y^{k+\ell-\beta-j}\right|_{y=1} = \binom{k-j+\ell-\beta}{\ell-\beta} \end{aligned} \tag{19}$$

So we can invert (18) term by term. We get

$$\begin{aligned} m(k,\theta,\ell,\phi) = & \frac{\rho(\phi)\rho(\theta)\left(2+\frac{\mu}{\beta}\right)\Gamma\left(\beta+2+\theta+\frac{\mu}{\beta}\right)}{(\beta-1+\theta)!(\beta-1+\phi)!} \\ & \times\sum_{j=\beta}^{\infty}\frac{\Gamma\left(2+\frac{\mu}{\beta}+k+\beta+\theta+\phi\right)}{\Gamma\left(k+2+\theta+\frac{\mu}{\beta}\right)}\binom{k-j+\ell-\beta}{\ell-\beta}. \end{aligned} \tag{20}$$

From this, we readily obtain

$$\begin{aligned} n(k,\theta,\ell,\phi) = & \rho(\phi)\rho(\theta)\frac{\Gamma\left(\beta+2+\theta+\frac{\mu}{\beta}\right)}{\Gamma\left(3+\frac{\mu}{\beta}+k+\ell+\theta+\phi\right)}\frac{(k-1+\theta)!(\ell-1+\phi)!}{(\beta-1+\theta)!(\beta-1+\phi)!} \\ & \times\left(2+\frac{\mu}{\beta}\right)\sum_{j=\beta}^{k}\frac{\Gamma\left(2+\frac{\mu}{\beta}+j+\beta+\theta+\phi\right)}{\Gamma\left(j+2+\theta+\frac{\mu}{\beta}\right)}\binom{k-j+\ell-\beta}{\ell-\beta}. \end{aligned} \tag{21}$$

The last step is to abridge this quantity and the desired NNQDD distribution, that is, $P(\ell,\phi|k,\theta)$. Remember that the NNQDD is the fraction of (ℓ,ϕ) nodes among the neighbors of a (k,θ) node. To obtain this fraction, we first need to obtain the total number of neighbors of (k,θ) nodes, then find the number of (ℓ,ϕ) nodes among these nodes, and divide the latter by the former. The total number of neighbors of (k,θ) nodes is simply $kNn(k,\theta)$.

The number of (ℓ,ϕ) nodes among them equals $\big[n(k,\theta,\ell,\phi)+n(\ell,\phi,k,\theta)\big]N$, because the (ℓ,ϕ) node can both be the parent or the child of the a (k,θ) node to be connected to it. So we have $P(\ell,\phi|k,\theta)=\frac{n(k,\theta,\ell,\phi)+n(\ell,\phi,k,\theta)}{kP(k,\theta)}$. Inserting the results of (21) and (1) into this expression and simplifying the results, we obtain

$$P(\ell,\phi|k,\theta)=\frac{\rho(\phi)}{k}\frac{\Gamma\left(k+\theta+3+\frac{\mu}{\beta}\right)}{\Gamma\left(k+\theta+3+\frac{\mu}{\beta}+\ell+\phi\right)}\frac{(\ell-1+\phi)!}{(\beta-1+\phi)!}\Gamma\left(\beta+2+\phi+\frac{\mu}{\beta}\right)\times$$
$$\left[\sum_{j=\beta+1}^{k}\frac{\Gamma\left(j+\theta+2+\frac{\mu}{\beta}+\beta+\phi\right)\binom{k-j+\ell-\beta}{\ell-\beta}}{\Gamma\left(j+\theta+2+\frac{\mu}{\beta}\right)\Gamma\left(\beta+2+\phi+\frac{\mu}{\beta}\right)}+\sum_{j=\beta+1}^{\ell}\frac{\Gamma\left(j+\theta+2+\frac{\mu}{\beta}+\beta+\phi\right)\binom{\ell-j+k-\beta}{k-\beta}}{\Gamma\left(j+\phi+2+\frac{\mu}{\beta}\right)\Gamma\left(\beta+2+\theta+\frac{\mu}{\beta}\right)}\right]. \tag{22}$$

REFERENCES

[1] S. L. Feld, "Why your friends have more friends than you do," *AJS*, vol. 96, no. 6, pp. 1464–77, 1991.

[2] W. Ezra and J. T. Zuckerman, "What makes you think youre so popular? self-evaluation maintenance and the subjective side of the friendship paradox," *Social Psychology Quarterly*, vol. 64, no. 3,207-223, 2001.

[3] M. Garcia-Herranz, E. Moro, M. Cebrian, N. Christakis, and J. Fowler, "Using friends as sensors to detect global-scale contagious outbreaks," *PLoS ONE*, vol. 9, no. 4, p. e92413, 04 2014.

[4] Y. Kryvasheyeu, H. Chen, E. Moro, P. Van Hentenryck, and M. Cebrian, "Performance of social network sensors during hurricane sandy," *arXiv preprint arXiv:1402.2482*, 2014.

[5] Y. H. Eom and H. H. Jo, "Generalized friendship paradox in complex networks: The case of scientific collaboration," *Sci. Rep.*, vol. 4, 2014.

[6] N. O. Hodas, F. Kooti, and K. Lerman, "Friendship paradox redux: Your friends are more interesting than you." *ICWSM*, vol. 13, pp. 8–10, 2013.

[7] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[8] G. Bianconi and A.-L. Barabási, "Competition and multiscaling in evolving networks," *EPL (Europhysics Letters)*, vol. 54, no. 4, p. 436, 2001.

[9] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, "Scale-free networks from varying vertex intrinsic fitness," *Physical review letters*, vol. 89, no. 25, p. 258702, 2002.

[10] V. D. Servedio, G. Caldarelli, and P. Butta, "Vertex intrinsic fitness: How to produce arbitrary scale-free networks," *Physical Review E*, vol. 70, no. 5, p. 056126, 2004.

[11] I. Smolyarenko, K. Hoppe, and G. Rodgers, "Network growth model with intrinsic vertex fitness," *Physical Review E*, vol. 88, no. 1, p. 012805, 2013.

[12] B. Fotouhi and M. Rabbat, *The European Physical Journal B*, vol. 86, no. 12, 2013.

[13] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, "Structure of growing networks with preferential linking," *Phys. Rev. Lett.*, vol. 85, pp. 4633–4636, 2000.

[14] P. L. Krapivsky and S. Redner, "Organization of growing random networks," *Physical Review E*, vol. 63, no. 6, p. 066123, 2001.

CHAPTER 10

Paper: Quantifying and Measuring the Friendship Paradox

The material presented in this chapter was published in the following conference proceedings:

N. Momeni, M. Rabbat, "Measuring the Generalized Friendship Paradox in Networks with Quality-Dependent Connectivity", Complex Networks VI. Springer International Publishing, 2015. 45-55.

Please note that the references of the manuscript are listed at the end of this chapter.

Measuring the Generalized Friendship Paradox in Networks with Quality-dependent Connectivity

Naghmeh Momeni and Michael G. Rabbat

Abstract

The friendship paradox is a sociological phenomenon stating that most people have fewer friends than their friends do. The generalized friendship paradox refers to the same observation for attributes other than degree, and it has been observed in Twitter and scientific collaboration networks. This paper takes an analytical approach to model this phenomenon. We consider a preferential attachment-like network growth mechanism governed by both node degrees and 'qualities'. We introduce measures to quantify paradoxes, and contrast the results obtained in our model to those obtained for an uncorrelated network, where the degrees and qualities of adjacent nodes are uncorrelated. We shed light on the effect of the distribution of node qualities on the friendship paradox. We consider both the mean and the median to measure paradoxes, and compare the results obtained by using these two statistics.

I. Introduction

The friendship paradox, introduced by Feld [1], is a sociological observation that says most people are less popular than their friends on average. It is called a 'paradox' because, while most people believe that they are more popular than their friends [2], Feld observed that the converse is actually true. There are more recent observations agreeing with Felds', that study online environments. For example on Twitter, people you follow and also your followers have, on average, more followers than you do. They also follow more people than you do [3]. On Facebook, your friends have, on average, more friends than you do [4].

The friendship paradox is about the inter-nodal inequality of the degrees. What happens if we consider other attributes? This is the focus of the 'Generalized Friendship Paradox' [5], [6]. For example on Twitter, your friends on average tweet more and also share more viral content than you [3], [7]. In the scientific collaboration networks your collaborators have on average more publications, more citations and more collaborators than you do [5].

The friendship paradox has applications in spotting influential nodes. In [8], it is used for finding high-degree nodes for efficient vaccination. In order to sample a node with above average degree, a node is chosen uniformly at random and one of their neighbours will be sampled. In [9], the friendship paradox is used for the early detection of flu outbreaks among college students. In [10], it is utilized to derive early-warning sensors during catastrophic events such as hurricanes.

In this paper, first we explain a quality-dependent preferential attachment scheme introduced in [11]. Then, we introduce measures to quantify the mean and the median paradoxes. In Section 4 these measures are computed numerically on the networks generated with the quality-dependent model and also uncorrelated networks. We compare the results obtained in these networks using both the mean and the median statistics. Furthermore, we study the effect of node quality distribution on the quality and friendship paradoxes.

II. Model, Notation and Terminology

We consider a *quality-based preferential attachment* (QPA) model, identical to the model proposed and analysed in [11]. It is similar to the Barabasi-Albert model [12], but incorporates node qualities. Each incoming node has β links, and a discrete quality θ drawn from a distribution $\rho(\theta)$ that is assigned to it upon birth. The probability of an existing node x with degree k_x and quality θ_x (at the instant) receiving a new link is proportional to $k_x + \theta_x$.

Once assigned, the quality of a node does not change. We denote the mean of the quality distribution by μ. Following [11], as the number of nodes tends to infinity, $P(k, \theta)$, the

fraction of nodes with degree k and quality θ is given by:

$$P(k,\theta) = \rho(\theta)\big(2+\frac{\mu}{\beta}\big)\frac{\Gamma(k+\theta)}{\Gamma(\beta+\theta)}\frac{\Gamma\big(\beta+\theta+2+\frac{\mu}{\beta}\big)}{\Gamma\big(k+\theta+3+\frac{\mu}{\beta}\big)}u(k-\beta). \tag{1}$$

In [11] the nearest-neighbor distribution, i.e., the fraction of neighbors of a node with degree k and quality θ who has degree ℓ and quality ϕ is given by:

$$P(\ell,\phi|k,\theta) = \frac{\rho(\phi)}{k}\frac{\Gamma\left(k+\theta+3+\frac{\mu}{\beta}\right)}{\Gamma\left(k+\theta+3+\frac{\mu}{\beta}+\ell+\phi\right)}\frac{(\ell-1+\phi)!}{(\beta-1+\phi)!}\Gamma\left(\beta+2+\phi+\frac{\mu}{\beta}\right)\times$$
$$\left[\sum_{j=\beta+1}^{k}\frac{\Gamma\left(j+\theta+2+\frac{\mu}{\beta}+\beta+\phi\right)\binom{k-j+\ell-\beta}{\ell-\beta}}{\Gamma\left(j+\theta+2+\frac{\mu}{\beta}\right)\Gamma\left(\beta+2+\phi+\frac{\mu}{\beta}\right)}+\sum_{j=\beta+1}^{\ell}\frac{\Gamma\left(j+\theta+2+\frac{\mu}{\beta}+\beta+\phi\right)\binom{\ell-j+k-\beta}{k-\beta}}{\Gamma\left(j+\phi+2+\frac{\mu}{\beta}\right)\Gamma\left(\beta+2+\theta+\frac{\mu}{\beta}\right)}\right]. \tag{2}$$

III. Measures of Friendship and Quality Paradoxes

By marginalizing the joint distribution $P(k,\theta)$ we can find the degree distribution, denoted by $P(k)$. Also, from the nearest-neighbor distribution (2), we can find the expected value of the qualities of neighbors of a node with quality θ and also the expected value of the degrees of neighbors of a node with degree k. This allows us to investigate when the quality paradox (hereinafter QP) and the friendship paradox (hereinafter FP) are in force, and which nodes in the network exhibit the paradox.

Let us also define the 'median' version of the paradoxes, following [7]. In the median version, instead of the average values of quality or degree of neighbors, we use the median values. A node experiences the median QP (FP), if its quality (degree) is less than the quality (degree) of at least half of its neighbors.

Throughout the paper, the superscript NN denotes Nearest-Neighbor. Let us denote the median operator by $M\{\cdot\}$. For example, $M\{\phi^{NN}|\theta\}$ denotes the median value of ϕ under the distribution $P(\phi|\theta)$, and is a function of θ. Also note that every measure we introduce here is by nature a function of the parameters of the quality distribution. For example, if the exponential decay quality distribution is considered, the measures will depend on the decay factor. We denote the parameter of the quality distribution by x. Using this notation,

we define the critical values for the mean and the median paradoxes as follows:

$$\text{mean: } \begin{cases} \widetilde{\theta}_c(x) \stackrel{\text{def}}{=} \max\left\{\theta \middle| \theta < E\{\phi^{\text{NN}}|\theta\}\right\} \\ \widetilde{k}_c(x) \stackrel{\text{def}}{=} \max\left\{k \middle| k < E\{\ell^{\text{NN}}|k\}\right\} \end{cases} \text{, median: } \begin{cases} \hat{\theta}_c(x) \stackrel{\text{def}}{=} \max\left\{\theta \middle| \theta < M\{\phi^{\text{NN}}|\theta\}\right\} \\ \hat{k}_c(x) \stackrel{\text{def}}{=} \max\left\{k \middle| k < M\{\ell^{\text{NN}}|k\}\right\} \end{cases} . \quad (3)$$

In other words, $\widetilde{\theta}_c(x)$ is the highest quality that a node can have, given that its quality is lower than the average quality of its neighbors. Similarly, $\widetilde{k}_c(x)$ is the highest degree that a node can have, given that it exhibits the mean FP. For the median version of the paradox, we have $\hat{\theta}_c(x)$ and $\hat{k}_c(x)$. So $\hat{\theta}_c(x)$ is the highest quality that a node exhibiting the median QP can have. Let us also emphasize that we use the following convention with regards to the median throughout the paper: the median of the probability distribution $g(x)$ (with CDF $G(x)$) is the minimum value of x for which $G(x) \geq \frac{1}{2}$. For example, for $g(x) = \frac{1}{2}\delta[x] + \frac{1}{2}\delta[x-5]$, the median is $x = 0$.

We now define similar quantities for an 'uncorrelated network'. In this network the qualities are assigned to nodes in an identical way to the QPA model, but the attachment of new nodes to existing nodes depends on neither the degrees nor the qualities of the existing nodes. In this network the properties of a node are uncorrelated with the properties of its neighbors. We denote this case by superscript u. For this network we have $P^u(\ell, \phi|k, \theta) = P(\ell, \phi)$ and $P^u(\phi|\theta) = \rho(\phi)$. For the critical values of the mean and the median paradoxes, we have:

$$\begin{cases} \widetilde{\theta}^u_c(x) \stackrel{\text{def}}{=} \max\left\{\theta \middle| \theta < E\{\phi^{\text{NN}}|\theta\}\right\} = \max\left\{\theta \middle| \theta < \underbrace{E\{\phi\}}_{=\mu}\right\} = \mu(x) - 1 \\ \hat{\theta}^u_c(x) \stackrel{\text{def}}{=} \max\left\{\theta \middle| \theta < M\{\phi^{\text{NN}}|\theta\}\right\} = \max\left\{\theta \middle| \theta < \underbrace{M\{\phi\}}_{=\hat{\theta}}\right\} = \hat{\theta}(x) - 1 \end{cases} . \quad (4)$$

Similarly, for degrees we have: $\widetilde{k}^u_c(x) = \overline{k}(x) - 1$ and $\hat{k}^u_c(x) = \hat{k}(x) - 1$.

We are also interested in the fraction of all nodes that experience each type of paradoxes. This is equal to the fraction of nodes with their attribute below the corresponding critical

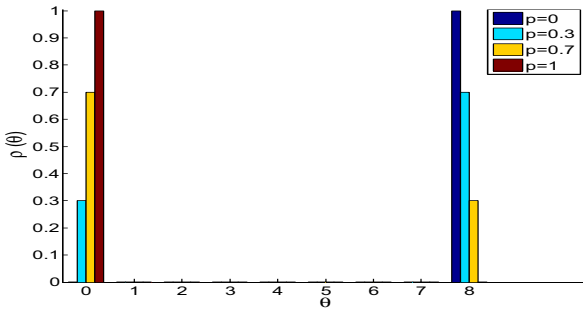
value. We denote these quantities by:

$$\text{mean:} \begin{cases} \widetilde{F}_\theta(x) = \sum_{\theta \le \bar{\theta}_c(x)} \rho(\theta) \\ \widetilde{F}_k(x) = \sum_{k \le \bar{k}_c(x)} P(k) \end{cases}, \qquad \text{median:} \begin{cases} \hat{F}_\theta(x) = \sum_{\theta \le \hat{\theta}_c(x)} \rho(\theta) \\ \hat{F}_k(x) = \sum_{k \le \hat{k}_c(x)} P(k) \end{cases}. \tag{5}$$

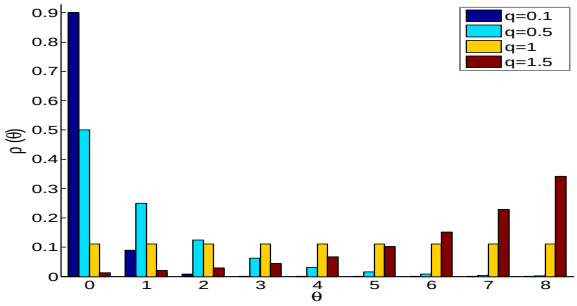
IV. Results and Discussion

In this paper we consider two quality distributions for expository purposes. The first one is the Bernoulli distribution, where nodes have quality 0 (with probability p) or quality $\theta_{\max}$ (with probability $1-p$). The other one is the discrete exponential distribution, with decay factor q. The probability of quality θ is proportional to q^θ, and the maximum value of θ is denoted by $\theta_{\max}$. Figure 1 depicts these quality distributions for four example values of p and q. Note that for $q < 1$, the exponential distribution is a decreasing function of quality and $\mu > \hat{\theta}$, and for $q > 1$, the distribution is increasing function of quality and $\mu < \hat{\theta}$. Also for the Bernoulli distribution note that, with the convention we use for the median, the value of the median is zero if $p \geq \frac{1}{2}$, and the median is equal to $\theta_{\max}$ if $p < \frac{1}{2}$. For each distribution, we have numerically computed all the introduced measures for four different values of β and four different values of $\theta_{\max}$.

Critical values obtained for two distributions are presented in Figure 2. These values are computed using the closed form expressions mentioned in Section 2. From Figure 2a we can learn about the differences between the networks that the QPA model generates and an uncorrelated network. In an uncorrelated network the probabilities of a random node being connected to a neighbor with quality 0 and $\theta_{\max}$ are equal to p and $1-p$, respectively (regardless of the quality of the node). If the majority of the neighbors have quality zero ($p \geq 0.5$), the median is zero. Similarly, if the majority have quality $\theta_{\max}$ ($p < 0.5$), the median is $\theta_{\max}$. So if $p < 0.5$, nodes with qualities up to $\theta_{\max} - 1$ experience the median QP and $\hat{\theta}_c^u = \theta_{\max} - 1$. Conversely, if $p \geq 0.5$, $\hat{\theta}_c^u = 0$. This explains the abrupt drop in $\hat{\theta}_c^u$ in Figure 2a. On the other hand, in the QPA model, this transition takes place at a p greater than 0.5. This means that upto some point beyond $p = 0.5$, although the probability of $\theta = 0$ is higher than that of $\theta = \theta_{\max}$, the majority of the friends of each node have



(a) Bernoulli distribution with $p = 0, 0.3, 0.7, 0.1$. The cases of $p = 0$ and $p = 1$ correspond to conventional Barabasi-Albert and shifted-linear preferential attachment networks, respectively.

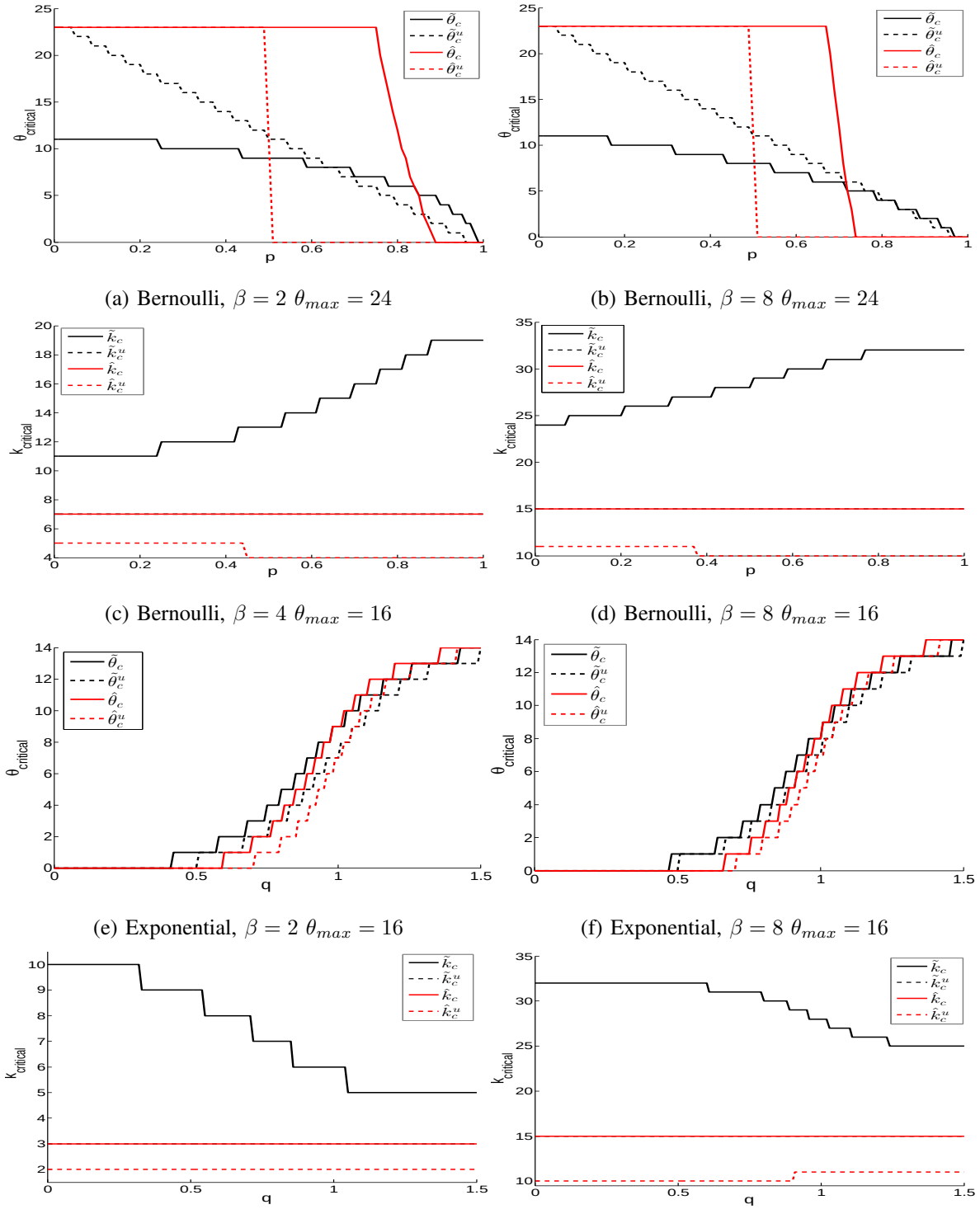


(b) Exponential distribution for decay factor $q = 0.1, 0.5, 1, 1.5$. The special case of $q = 1$ corresponds to a uniform distribution supported in the interval $0 \leq \theta \leq \theta_{\max}$.

Fig. 1: Examples of the quality distributions used in this paper with $\theta_{\max} = 8$. Four instances of each type is depicted.

quality $\theta_{\max}$. There is a region for $p > 0.5$, where the majority of the network have quality zero, but the majority of the neighbors of most nodes have quality $\theta_{\max}$. This indicates quality disassortativity, since low quality nodes are mostly connected to nodes with high qualities.

For the mean version of the QP, we consider the example case of $p = 0.2$ for discussion. In an uncorrelated network, each node (with any quality) is connected to neighbors with quality 0 and $\theta_{\max}$ with probabilities 0.2 and 0.8, respectively. So the average of the qualities of its neighbors is $0.8\,\theta_{\max}$. So nodes with quality less than $0.8\,\theta_{\max}$ experience the mean QP. On the other hand, in the QPA model $\widetilde{\theta}_c < \widetilde{\theta}_c^u$ at $p = 0.2$. This means that nodes whose qualities are between $\widetilde{\theta}_c$ and $\widetilde{\theta}_c^u$, do not experience the mean QP in the QPA model (while they do experience this paradox in the uncorrelated case). We deduce that these nodes are connected to quality zero nodes with a higher probability than 0.2. This reduces the average quality of their niehgbors. Now consider the example case of $p = 0.8$. In this case, $\widetilde{\theta}_c^u < \widetilde{\theta}_c$. This means that nodes with quality between $\widetilde{\theta}_c$ and $\widetilde{\theta}_c^u$ experience the mean QP in the proposed model, while they do not experience it in the uncorrelated



(a) Bernoulli, $\beta = 2$ $\theta_{max} = 24$

(b) Bernoulli, $\beta = 8$ $\theta_{max} = 24$

(c) Bernoulli, $\beta = 4$ $\theta_{max} = 16$

(d) Bernoulli, $\beta = 8$ $\theta_{max} = 16$

(e) Exponential, $\beta = 2$ $\theta_{max} = 16$

(f) Exponential, $\beta = 8$ $\theta_{max} = 16$

(g) Exponential, $\beta = 2$ $\theta_{max} = 16$

(h) Exponential, $\beta = 8$ $\theta_{max} = 16$

Fig. 2: Critical values for quality and degree as defined in (3) and (4) computed for Bernoulli and exponential quality distributions.

case. In an uncorrelated network these nodes would be connected to zero and $\theta_{\max}$ quality nodes probabilities 0.8 and 0.2, respectively. However, in the QPA model, these nodes are connected to nodes with quality $\theta_{\max}$ with a probability higher than 0.2, and this increases the average quality of their neighbors, making them subject to the mean QP.

Comparing Figure 2b with 2a we observe the curves are similar, but the difference between the QPA model and the uncorrelated case is smaller in Figure 2b. For example, the drop in the $\hat{\theta}_c$ curve is closer to the drop in $\hat{\theta}_c^u$ for the uncorrelated case. We conclude that increasing β decreases the difference between the QPA model and the uncorrelated case.

In Figure 2c, critical degrees are depicted. It can be observed that as p increases, $\widetilde{k}_c$ increases. Comparing Figures 2c and 2d, we observe that all the critical degrees are greater in the case of $\beta = 8$ than $\beta = 4$. Also the range of node degrees experiencing any type of paradoxes is wider in the $\beta = 8$ case.

From Figure 2e, we observe that for fixed decay factor, $\widetilde{\theta}_c \geq \widetilde{\theta}_c^u$ and $\hat{\theta}_c \geq \hat{\theta}_c^u$. This means that there exist values of θ that in the uncorrelated network experience QP, but in the proposed model they do not. So the range of possible values of quality that experience the QP is wider in the QPA model than in uncorrelated networks. This argument holds for both mean and median paradoxes.

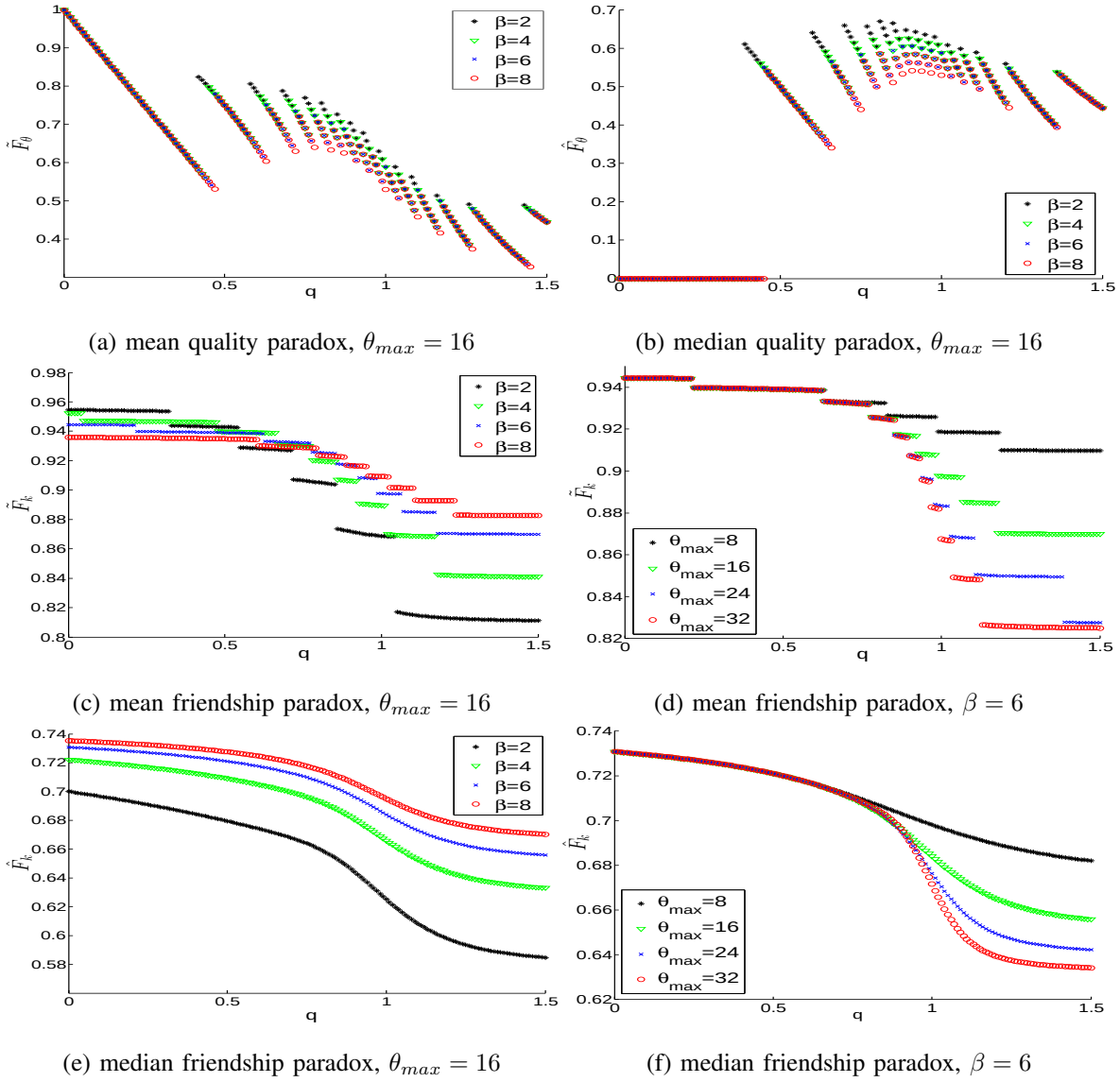
We also observe from Figure 2e that for $q < 1$, $\widetilde{\theta}_c \geq \hat{\theta}_c$ and $\widetilde{\theta}_c^u \geq \hat{\theta}_c^u$. Both of these inequalities flip in the case of $q > 1$. The main cause of this change of regime is the difference between the shape of the quality distribution for $q > 1$ and $q < 1$. When $q < 1$, the median paradox is stronger (using the terminology of [7]), that is, the median paradox applies to a smaller range of qualities than the mean paradox (for both the uncorrelated network and the QPA model). However, when $q > 1$, the median of the distribution is greater than the mean. As it can be observed in Figure 2e, there are values of θ that are subject to the median version of the paradox, but not to the mean version. This means that the term 'strong paradox' introduced in [7] is not applicable to this case, because the mean version provides a tighter range of qualities in paradox, as compared to the median version.

Another observable trend in Figure 2e is that the critical values of quality are a non-decreasing functions of q. This can be intuitively explained as follows. When q is low, the majority of the network is constituted by low quality nodes. The majority of the neighbors of a low quality node will also have low quality. So the node does not experience the paradox with high probability. When q increases, the number of nodes with higher quality increases, and a low quality node has a higher probability of being connected to those high quality nodes, which gives it a higher probability of experiencing paradox. Comparing Figure 2f with Figure 2e, we observe that as β varies $\widetilde{\theta}_c^u$ and $\hat{\theta}_c^u$ do not change, while the critical values of the QPA model get closer to those of the uncorrelated case. These figures only depict the results for two values of β, due to space limitations. The trend holds for the omitted figures. We conclude that *as β gets larger, the correlation of the quality of a node with the quality of its neighbors diminishes.*

In Figure 2g, the critical degrees (as defined in (3) and (4)) are depicted. It can be observed that as q increases, $\widetilde{k}_c$ decreases. Comparing Figures 2g and 2h, we observe that all the critical degrees are greater in the case of $\beta = 8$ than $\beta = 2$. Also the range of the degrees who experience paradox (of any type) is wider when $\beta = 8$. In both figures, we observe that the mean FP is more sensitive to changes in the quality distribution than the median FP.

Figure 3 depicts the fraction of nodes in the quality and friendship paradoxes (as defined in (5)) when quality distribution is exponential. From Figure 3a we observe that, as q increases in the vicinity of zero, $\tilde{F}_\theta$, the fraction of nodes experiencing the mean QP (with qualities lower than $\widetilde{\theta}_c$) decreases, because increasing q increases the fraction of nodes with high qualities. The fraction $\tilde{F}_\theta$ has discontinuities at the values of q at which $\widetilde{\theta}_c$ is incremented by one. So all the nodes whose qualities where equal to the new $\widetilde{\theta}_c$ are taken into account as those who experience the mean QP, hence the abrupt jump.

The fraction of nodes in the median QP is depicted in Figure 3b. It can be seen that $\hat{F}_\theta$ has a similar behavior to that of $\tilde{F}_\theta$. Each discontinuity pertains to a value of q at which $\hat{\theta}_c$ increments. The main difference between Figures 3a and 3b is the behavior near $q = 0$. In the mean QP, when almost all nodes have quality zero, even one non-zero quality neighbor



(a) mean quality paradox, $\theta_{max} = 16$

(b) median quality paradox, $\theta_{max} = 16$

(c) mean friendship paradox, $\theta_{max} = 16$

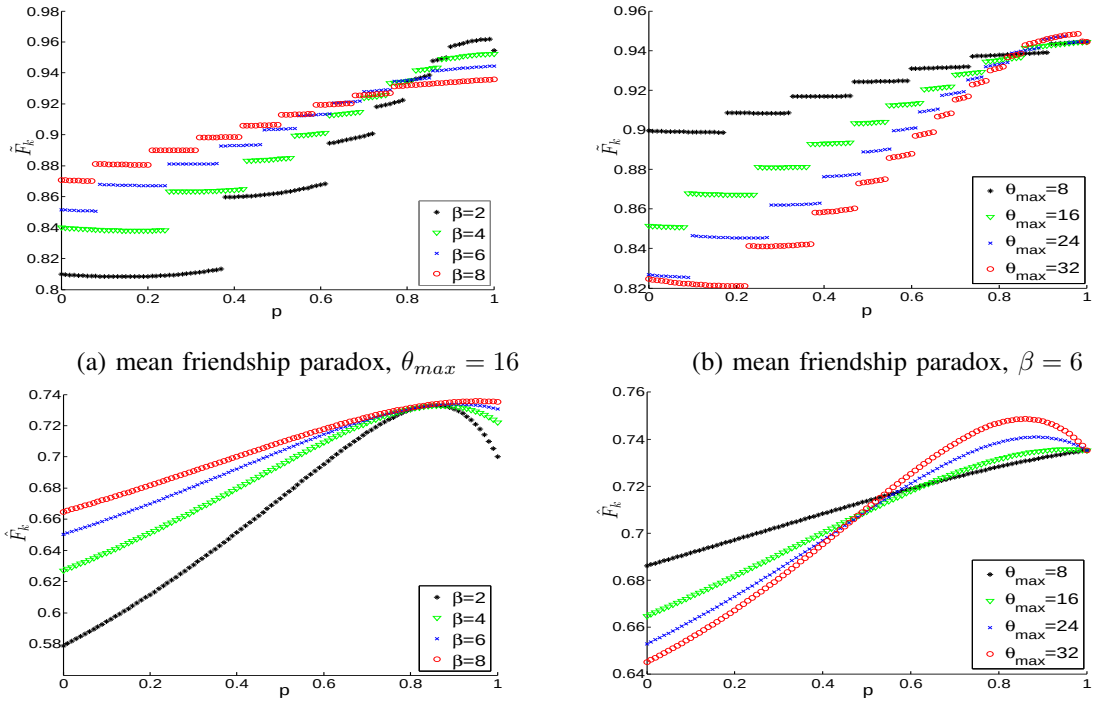
(d) mean friendship paradox, $\beta = 6$

(e) median friendship paradox, $\theta_{max} = 16$

(f) median friendship paradox, $\beta = 6$

Fig. 3: The fraction of nodes in the quality and friendship paradoxes when the quality distribution $\rho(\theta)$ is exponential.

elevates the average above zero, so all those zero-quality nodes experience the mean QP. However, in the median version, at least half of the friends of a zero-quality node must have non-zero quality. Also observe that for $q < 1$, we have $\widetilde{F}_\theta \geq \hat{F}_\theta$, i.e, the fraction of nodes in the mean QP is higher than the fraction of nodes in the median QP. But, for $q > 1$



(a) mean friendship paradox, $\theta_{max} = 16$

(b) mean friendship paradox, $\beta = 6$

(c) median friendship paradox, $\theta_{max} = 16$

(d) median friendship paradox, $\beta = 8$

Fig. 4: The fraction of nodes in the friendship paradox when the node quality distribution $\rho(\theta)$ is Bernoulli.

the inequality changes sides.

In Figures 3c and 3d, it can be observed that *for all values of* β *and* $\theta_{\max}$, *the majority of the nodes (over* 80%*) experience the mean FP.* Also, as q increases, $\widetilde{F}_k$ decreases. It means that the quality distribution affects the FP that depends solely on degrees. Through the quality-dependant network growth mechanism, the degree distribution, and hence the conditions under which a node experiences the FP, depend on the quality distribution. Also, it is observed in Figure 3c that as β increases, the sensitivity of $\widetilde{F}_k$ to variations of q decreases. This means that as the initial degree of nodes increases, the effect of the quality distribution on the FP diminishes. Because as β increases the final degrees of nodes increase, and for larger degrees $k + \theta$ is dominated by k; varying θ has less of an effect. Conversely, in Figure 3d, as $\theta_{\max}$ increases, the sensitivity of $\widetilde{F}_k$ to variations of q increases.

As the range of possible qualities becomes wider, the probability of having high values of θ that have significant roles in $k+\theta$ increases.

In Figures 3e and 3f, we observe that as q increases, $\hat{F}_k$ (the fraction of nodes experiencing the median FP) decreases. This is similar to the trend observed for $\widetilde{F}_k$ in Figures 3c and 3d. From Figure 3e we observe that $\hat{F}_k$ increases as β increases. From Figure 3f we observe that for a range of decay factors (up to around $q = 0.7$), $\theta_{\max}$ does not have a significant effect on $\hat{F}_k$, but beyond that point, $\hat{F}_k$ decreases as $\theta_{\max}$ increases. Also, comparing Figures 3e and 3f with Figures 3c and 3d, we assert that $\hat{F}_k \leq \widetilde{F}_k$. In other words, *the median FP is always stronger than the mean FP, regardless of the quality distribution*.

The fraction of nodes experiencing the FP when the quality distribution is Bernoulli are depicted in Figure 4. From Figures 4a and 4b we observe that as p increases, $\widetilde{F}_k$ (the fraction of nodes experiencing the mean FP) increases. From Figure 4a we deduce that as β increases, the sensitivity of $\widetilde{F}_k$ to variations of p decreases. Also, in Figure 4b it is observed that as $\theta_{\max}$ increases, the sensitivity of $\widetilde{F}_k$ to variations of p increases (similar to Figures 3c and 3d).

From Figure 4c we observe that as β increases, $\hat{F}_k$ (the fraction of nodes experiencing the median FP) increases. From Figure 4d we observe that as $\theta_{\max}$ increases, the sensitivity of $\hat{F}_k$ to the variations of p increases. Comparing Figures 4a and 4b with Figures 4c and 4d we deduce that for each value of p, we have $\hat{F}_k \leq \widetilde{F}_k$, i.e., *the fraction of nodes experiencing the mean FP is higher than nodes in the median FP regardless of the quality distribution*.

V. Summary and Future Work

In this paper we studied the friendship and the generalized friendship paradoxes on networks grown under a quality-based preferential attachment scheme. To this end, we introduced measures, such as quality and degree critical values, and fraction of nodes that experience each paradox. In each case, we considered the mean and the median to characterize the paradox. We compared the results to the uncorrelated network where the qualities and degrees of neighbors are uncorrelated. We considered Bernoulli and

exponential distributions for qualities.

For the exponential quality distribution, the critical quality of the uncorrelated case is always smaller than that of the QPA model. This means that the range of possible values of the quality that experience paradox is wider in the QPA model than in the uncorrelated case. We also observed that as β increases, the nearest-neighbor quality correlation decreases. In other words, the critical values of the proposed model converge to those of the uncorrelated case. For the exponential quality distribution we also observe that when $q < 1$ (which makes the median smaller than the mean), the median QP is stronger than the mean QP for both the QPA model and the uncorrelated case. The converse is true for $q > 1$. For all values of β, $\theta_{\max}$, over 80% of nodes experience the mean FP. We observed that changing the distribution of qualities affects the FP (in addition to the QP). This effect is strengthened when β decreases or when $\theta_{\max}$ increases. Also, it was observed that regardless of the quality distribution, the median FP is always stronger than the mean FP.

Plausible extensions of the present contribution are as follows. We can apply the measures introduced here to real networks, and compare the results, and also compare them with networks synthesized with arbitrary quality distributions. This enables us to investigate what type of quality distribution best characterizes a given network.

REFERENCES

[1] S. L. Feld, "Why your friends have more friends than you do," *American Journal of Sociology*, vol. 96, no. 6, pp. 1464–77, 1991.

[2] W. Ezra and J. T. Zuckerman, "What makes you think youre so popular? self-evaluation maintenance and the subjective side of the friendship paradox," *Social Psychology Quarterly*, vol. 64, no. 3,207-223, 2001.

[3] N. O. Hodas, F. Kooti, and K. Lerman, "Friendship paradox redux: Your friends are more interesting than you." *ICWSM*, vol. 13, pp. 8–10, 2013.

[4] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *arXiv preprint arXiv:1111.4503*, 2011.

[5] Y. H. Eom and H. H. Jo, "Generalized friendship paradox in complex networks: The case of scientific collaboration," *Sci. Rep.*, vol. 4, 2014.

[6] H. H. Jo and Y. H. Eom, "Generalized friendship paradox in networks with tunable degree-attribute correlation," *Phys. Rev. E*, vol. 90, p. 022809, 2014.

[7] F. Kooti, N. O. Hodas, and K. Lerman, "Network weirdness: Exploring the origins of network paradoxes," *arXiv preprint arXiv:1403.7242*, 2014.

[8] R. Cohen, S. Havlin, and D. Ben-Avraham, "Efficient immunization strategies for computer networks and populations," *Physical review letters*, vol. 91, no. 24, p. 247901, 2003.

[9] M. Garcia-Herranz, E. Moro, M. Cebrian, N. Christakis, and J. Fowler, "Using friends as sensors to detect global-scale contagious outbreaks," *PLoS ONE*, vol. 9, no. 4, p. e92413, 04 2014.

[10] Y. Kryvasheyeu, H. Chen, E. Moro, P. Van Hentenryck, and M. Cebrian, "Performance of social network sensors during hurricane sandy," *arXiv preprint arXiv:1402.2482*, 2014.

[11] B. Fotouhi, N. Momeni, and M. G. Rabbat, "Generalized friendship paradox: An analytical approach," in *International Conference on Social Informatics*. Springer, 2014, pp. 339–352.

[12] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

CHAPTER 11

Future Work

11.1 Multiplex Data

In many social network studies, multiple name generators are used and each respondent gives several distinct alter lists corresponding to different questions [PSC15]. Instead of a simple network, we can picture the system as a multi-layer network, where the same set of nodes have distinct sets of links between them. We saw a simplified case in Chapter 3, where the response mechanisms for the two layers were different (full response for the strong layer and FCD for the weak layer). In many contexts, FCD is employed on every layer. What would be needed in those settings is an inference framework for given M number of layers and $B_1 \ldots, B_M$ cutoffs. Like the case considered in Chapter 3, the information on each layer will play a role in the estimations of other layers.

11.2 Additional Socio-centric Data

The sampling model considered in this thesis is designed to emulate the convention: in most cases, we interview a number of respondents and the survey process ends there. But in rare cases, investigators succeed at tracking the mentioned alters and interview them. This sociocentric approach obviously provides richer data. Incorporating the

second-wave data into the estimation is an open problem. We can use the additional information to improve the estimates, as well as providing estimates for new parameters that one-wave data is unable to capture (such as neighbor average degree, average path length, and higher degree moments). Moreover, we can assess the quality of the first-wave responses by looking at the rate of reciprocated ties. That is, a fraction of alters who were mentioned by first-wave egos will not mention those egos as their own friends [BN13]. This can be helpful for studying and characterizing patterns of reciprocity, and iteratively, we can use it to reduce the error in the second-wave data.

11.3 Heterogeneous Tie Strength

In Chapter 3, we assumed for simplicity that strong and weak ties were treated via distinct survey questions, and that they could be distinguished in the data. There is evidence in evolutionary psychology and neuroscience that the brain of Sapiens is evolved to retain social connections with certain capacities. Our social ties are characterized by distinct layers of intimacy [ZSHD05, HBD08, SDBA12, Dun14]. The innermost layer of our personal social network comprises on average five members, and consists of people we would seek personal advice from, share secrets with, and seek help from in times of serious emotional distress or financial trouble. The next layer has on average close to 15 members. It is called the sympathy group, and consists of people we regularly see. This group provides instrumental support and coalitions. The next layer has on average around 150 members. It comprises people with whom we have a bilateral relationship and is governed by norms of reciprocity. The next layer has on average around 500 members, and include people we count as our 'acquaintances'. The last layer has on average around 1500 members and includes people whose faces we can identify and we know by name.

Interestingly, repeating the analyses of these layers for online social networks, researchers found one primary layer with on average 1.5 members (the authors of [DACP15] speculate that it might be because, as for example is observed in [SLL$^+$14], for most men this layer comprises a partner, and for most women it includes their partner and a closest friend). These discrete layers of tie intensity (as opposed to continuous) with constant ratio can be used to model strength of social ties more realistically, and using a parametric model, we can use maximum likelihood methods to obtain estimates for

the ratio. This ultimately provides a weighted network, but once the parameters of the intensity distribution is estimated, many other properties can be derived from this distribution.

11.4 Effect of Inequalities in Social Networks on Subjective Well-being

Inequality is the most central question in sociology and one of the most central questions in economics. In this thesis, we showed that persistent patters of local inequalities can exist across the network both in terms of both structural and non-structural properties. There is overwhelming evidence in research on happiness in psychology and economics that points to the 'comparative' nature of happiness. That is, an important component that drives personal happiness is comparing own standing relative to others in society and our social group [Eas01, BC80, Vee91, CO96, DL00, FiC05, Lut05, GF06, KK07, BGOQ08, HH08, CWNK09, BBM10, AKGK12, KPMD$^+$12, KBRN14] . Our findings in this thesis can have significant impact on the theory of happiness. Consistent local inequalities might be a consequence of the network nature of social life. That is, the vast majority of people might be locally worse off when they compare themselves to their network neighbors, and this can have a share in the unhappiness of almost everyone. We do not mean that this makes everybody unhappy. Personalities differ and the effect of relative comparisons differ in different people. We mean that this network-inequality component *exists* for almost everybody. Its strength might differ from person to person. This would be a new insight in happiness research.

11.5 Interplay between Structural and Non-structural Inequality

A much-studied topic in sociology is that how the rich also have access to better social capital. That is, being rich does not only mean having more wealth, but is also usually associated with having ties to more influential people in different contexts [Lin00, Bou11, Nar02, Woo01, Ros00, Cle05, HDT$^+$07]. On the other hand, economic experiments have shown that network structure plays a key role in determining collective cooperative outcomes [RAC11, RNFC14, NSRC15] ('cooperation' in

these experiments is usually studied in game-theoretical settings, such as the Prisoner's Dilemma or the Public Goods Game). Experiments have also shown that visibility of wealth decreases levels of cooperation, and also intensifies inequality. What has not been investigated, though, is the direct effect of structural inequalities on non-structural inequalities. That is, for example, in the same setup as in [NSRC15], instead of initially-unequal endowments, have equal endowments, and investigate how the accumulation of wealth relates to the network position of individuals. In socially-unrealistic settings such as that of [NSRC15], the underlying network is considered to be Erdős-Rényi. From previous discussions we know that these networks have low structural inequalities, unlike real social networks. We propose employing highly-unequal network structures, and then studying how the widespread structural inequalities translate into differences in wealth as time passes. This would shed new light and provide invaluable insight into the origins of social inequalities.

REFERENCES

[AB14] John R Anderson and Gordon H Bower. *Human associative memory*. Psychology press, 2014.

[ABD+14] Fabrizio Altarelli, Alfredo Braunstein, Luca DallâĂŹAsta, Alejandro Lage-Castellanos, and Riccardo Zecchina. Bayesian inference of epidemics on networks via belief propagation. *Physical review letters*, 112(11):118701, 2014.

[ACKM09] Dimitris Achlioptas, Aaron Clauset, David Kempe, and Cristopher Moore. On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs. *Journal of the ACM (JACM)*, 56(4):21, 2009.

[AFLŠ+15] Nino Antulov-Fantulin, Alen Lančić, Tomislav Šmuc, Hrvoje Štefančić, and Mile Šikić. Identification of patient zero in static and temporal networks: Robustness and limitations. *Physical review letters*, 114(24):248701, 2015.

[AJC64] C Norman Alexander Jr and Ernest Q Campbell. Peer influences on adolescent educational aspirations and attainments. *American Sociological Review*, pages 568–575, 1964.

[AKB+95] Mark D Alicke, Mary L Klotz, David L Breitenbecher, Tricia J Yurak, and Debbie S Vredenburg. Personal contact, individuation, and the better-than-average effect. *Journal of personality and social psychology*, 68(5):804, 1995.

[AKGK12] Cameron Anderson, Michael W Kraus, Adam D Galinsky, and Dacher Keltner. The local-ladder effect: Social status and subjective well-being. *Psychological Science*, 23(7):764–771, 2012.

[AKM+07] by Hunt Allcott, Dean Karlan, Markus M Möbius, Tanya S Rosenblat, and Adam Szeidl. Community size and network closure. *The American economic review*, pages 80–85, 2007.

[AM89] Blake E Ashforth and Fred Mael. Social identity theory and the organization. *Academy of management review*, 14(1):20–39, 1989.

[Bau05] Chris T Bauch. Imitation dynamics predict vaccinating behaviour. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1573):1669–1675, 2005.

[BBM10] Christopher J Boyce, Gordon DA Brown, and Simon C Moore. Money and happiness: Rank of income, not income, affects life satisfaction. *Psychological Science*, 21(4):471–475, 2010.

[BC80] Morty Bernstein and Faye Crosby. An empirical examination of relative deprivation theory. *Journal of Experimental Social Psychology*, 16(5):442–456, 1980.

[BCZ10] Jonathan D Bohlmann, Roger J Calantone, and Meng Zhao. The effects of market network heterogeneity on innovation diffusion: An agent-based modeling approach. *Journal of Product Innovation Management*, 27(5):741–760, 2010.

[BE04] Chris T Bauch and David JD Earn. Vaccination and the theory of games. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36):13391–13394, 2004.

[Ben11] Yochai Benkler. *The penguin and the leviathan: How cooperation triumphs over self-interest*. Crown Business, 2011.

[BFB91] Ann Bowling, Morag Farquhar, and Peter Browne. Life satisfaction and associations with social network and support variables in three samples of elderly people. *International journal of geriatric psychiatry*, 6(8):549–566, 1991.

[BGOQ08] Gordon DA Brown, Jonathan Gardner, Andrew J Oswald, and Jing Qian. Does wage rank affect employeesâĂŹ well-being? *Industrial Relations: A Journal of Economy and Society*, 47(3):355–389, 2008.

[BHH96] John J Beggs, Valerie A Haines, and Jeanne S Hurlbert. Revisiting the rural-urban contrast: Personal networks in nonmetropolitan and metropolitan settings1. *Rural sociology*, 61(2):306–325, 1996.

[BKW02] Jere R Behrman, Hans-Peter Kohler, and Susan Cotts Watkins. Social networks and changes in contraceptive use over time: Evidence from a longitudinal study in rural kenya. *Demography*, 39(4):713–738, 2002.

[BLA15] Fabrício Benevenuto, Alberto HF Laender, and Bruno L Alves. The h-index paradox: your coauthors have a higher h-index than you do. *Scientometrics*, pages 1–6, 2015.

[BN13] Brian Ball and Mark EJ Newman. Friendship networks and social status. *Network Science*, 1(01):16–30, 2013.

[Bou11] Pierre Bourdieu. The forms of capital.(1986). *Cultural theory: An anthology*, pages 81–93, 2011.

[Bur84] Ronald S Burt. Network items and the general social survey. *Social networks*, 6(4):293–339, 1984.

[Bur09] Ronald S Burt. *Structural holes: The social structure of competition*. Harvard university press, 2009.

[BVA05] Simona Bignami-Van Assche. Network stability in longitudinal data: A case study from rural malawi. *Social Networks*, 27(3):231–247, 2005.

[CF07] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[CF08] Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258, 2008.

[CF10] Nicholas A Christakis and James H Fowler. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9):e12948, 2010.

[CFC09] John T Cacioppo, James H Fowler, and Nicholas A Christakis. Alone in the crowd: the structure and spread of loneliness in a large social network. *Journal of personality and social psychology*, 97(6):977, 2009.

[CH56] Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277, 1956.

[CHBA03] Reuven Cohen, Shlomo Havlin, and Daniel Ben-Avraham. Efficient immunization strategies for computer networks and populations. *Physical review letters*, 91(24):247901, 2003.

[Chr10] Nicholas Christakis. The hidden influence of social networks. *TED: Ideas worth spreading*, 2010. retrieved at `http://www.ted.com/talks/nicholas_christakis_the_hidden_influence_of_social_networks/transcript?language=en`.

[CKM57] James Coleman, Elihu Katz, and Herbert Menzel. The diffusion of an innovation among physicians. *Sociometry*, pages 253–270, 1957.

[CL91] Karen E Campbell and Barrett A Lee. Name generators in surveys of personal networks. *Social networks*, 13(3):203–221, 1991.

[Cle05] Frances Cleaver. The inequality of social capital and the reproduction of chronic poverty. *World Development*, 33(6):893–906, 2005.

[CM05] Aaron Clauset and Cristopher Moore. Accuracy and scaling phenomena in internet mapping. *Physical Review Letters*, 94(1):018701, 2005.

[CNGP07] Kyung-Hee Choi, Zhen Ning, Steven E Gregorich, and Qi-chao Pan. The influence of social and sexual networks in the spread of hiv and syphilis among men who have sex with men in shanghai, china. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 45(1):77–84, 2007.

[CO96] Andrew E Clark and Andrew J Oswald. Satisfaction and comparison income. *Journal of public economics*, 61(3):359–381, 1996.

[Col88] James S Coleman. Social capital in the creation of human capital. *American journal of sociology*, pages S95–S120, 1988.

[CTS13] Emrah Cem, Mehmet Engin Tozal, and Kamil Sarac. Impact of sampling design in estimation of graph characteristics. In *Performance Computing and Communications Conference (IPCCC), 2013 IEEE 32nd International*, pages 1–10. IEEE, 2013.

[CTSMM02] Ivan D Chase, Craig Tovey, Debra Spangler-Martin, and Michael Manfredonia. Individual differences versus social dynamics in the formation of animal dominance hierarchies. *Proceedings of the National Academy of Sciences*, 99(8):5744–5749, 2002.

[CWNK09] Andrew E Clark, Niels Westergård-Nielsen, and Nicolai Kristensen. Economic satisfaction and income rank in small neighbourhoods. *Journal of the European Economic Association*, 7(2-3):519–527, 2009.

[CWY09] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[DACP15] Robin IM Dunbar, Valerio Arnaboldi, Marco Conti, and Andrea Passarella. The structure of online social networks mirrors those in the offline world. *Social Networks*, 43:39–47, 2015.

[DJBJ10] Sebastiano A Delre, Wander Jager, Tammo HA Bijmolt, and Marco A Janssen. Will it spread or not? the effects of social influences and network topology on innovation diffusion. *Journal of Product Innovation Management*, 27(2):267–282, 2010.

[DL00] Ed Diener and Richard E Lucas. Explaining differences in societal levels of happiness: Relative standards, need fulfillment, culture, and evaluation theory. *Journal of Happiness Studies*, 1(1):41–78, 2000.

[DR01] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[dSPK78] Ithiel de Sola Pool and Manfred Kochen. Contacts and influence. *Social networks*, 1(1):5–51, 1978.

[Dun14] ROBIN IM Dunbar. The social brain: Psychological underpinnings and implications for the structure of organizations. *Current Directions in Psychological Science*, 23(2):109–114, 2014.

[Eas01] Richard A Easterlin. Income and happiness: Towards a unified theory. *The economic journal*, 111(473):465–484, 2001.

[EJ14] Young-Ho Eom and Hang-Hyun Jo. Generalized friendship paradox in complex networks: The case of scientific collaboration. *Scientific reports*, 4, 2014.

[FBMG+07] Samuel R Friedman, Melissa Bolyard, Pedro Mateu-Gelabert, Paula Goltzman, Maria Pia Pawlowicz, Dhan Zunino Singh, Graciela Touze, Diana Rossi, Carey Maslow, Milagros Sandoval, et al. Some data-driven

reflections on priorities in aids network research. *AIDS and Behavior*, 11(5):641–651, 2007.

[FC+08] James H Fowler, Nicholas A Christakis, et al. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj*, 337:a2338, 2008.

[FCF17] Feng Fu, Nicholas A Christakis, and James H Fowler. Dueling biological and social contagions. *Scientific Reports*, 7, 2017.

[Fel91] Scott L Feld. Why your friends have more friends than you do. *American Journal of Sociology*, pages 1464–1477, 1991.

[FG77] Scott L Feld and Bernard Grofman. Variation in class size, the class size paradox, and some consequences for students. *Research in Higher Education*, 6(3):215–222, 1977.

[FiC05] Ada Ferrer-i Carbonell. Income and well-being: an empirical analysis of the comparison income effect. *Journal of Public Economics*, 89(5):997–1019, 2005.

[Fis82] Claude S Fischer. *To dwell among friends: Personal networks in town and city*. University of chicago Press, 1982.

[FMGC+07] Samuel R Friedman, Pedro Mateu-Gelabert, Richard Curtis, Carey Maslow, Melissa Bolyard, Milagros Sandoval, and Peter L Flom. Social capital or networks, negotiations, and norms? a neighborhood case study. *American journal of preventive medicine*, 32(6):S160–S170, 2007.

[Fra88] Ove Frank. Random sampling and social networks. a survey of various approaches. *Mathématiques et Sciences Humaines*, 104:19–33, 1988.

[FSC11] James H Fowler, Jaime E Settle, and Nicholas A Christakis. Correlated genotypes in friendship networks. *Proceedings of the National Academy of Sciences*, 108(5):1993–1997, 2011.

[Fuk01] Francis Fukuyama. Social capital, civil society and development. *Third world quarterly*, 22(1):7–20, 2001.

[GF06] Carol Graham and Andrew Felton. Inequality and happiness: insights from latin america. *Journal of Economic Inequality*, 4(1):107–122, 2006.

[GGA13] Clara Granell, Sergio Gómez, and Alex Arenas. Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Physical review letters*, 111(12):128701, 2013.

[GGLNT04] Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004.

[GHLH09] Jacob Goldenberg, Sangman Han, Donald R Lehmann, and Jae Weon Hong. The role of hubs in the adoption process. *Journal of marketing*, 73(2):1–13, 2009.

[GHMC+14] Manuel Garcia-Herranz, Esteban Moro, Manuel Cebrian, Nicholas A Christakis, and James H Fowler. Using friends as sensors to detect global-scale contagious outbreaks. *PloS one*, 9(4):e92413, 2014.

[GLA+07] Lazaros K Gallos, Fredrik Liljeros, Panos Argyrakis, Armin Bunde, and Shlomo Havlin. Improving immunization strategies. *Physical Review E*, 75(4):045104, 2007.

[GMT05] Ayalvadi Ganesh, Laurent Massoulié, and Don Towsley. The effect of network topology on the spread of epidemics. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 2, pages 1455–1466. IEEE, 2005.

[Gra73] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[Gre14] Joshua Greene. *Moral tribes: Emotion, reason, and the gap between us and them.* Penguin, 2014.

[Har] Frank Harary. Graph theory. 1969.

[HBD08] Russell A Hill, R Alexander Bentley, and Robin IM Dunbar. Network scaling reveals consistent fractal pattern in hierarchical mammalian societies. *Biology letters*, 4(6):748–751, 2008.

[HDT+07] Sam Hickey, Andries Du Toit, et al. Adverse incorporation, social exclusion and chronic poverty. *Series Editors*, page 134, 2007.

[Hei13] Fritz Heider. *The psychology of interpersonal relations.* Psychology Press, 2013.

[HH08] Ryan T Howell and Colleen J Howell. The relation of economic status to subjective well-being in developing countries: a meta-analysis. *Psychological bulletin*, 134(4):536, 2008.

[HK07] Stephane Helleringer and Hans-Peter Kohler. Sexual network structure and the spread of hiv in africa: evidence from likoma island, malawi. *Aids*, 21(17):2323–2332, 2007.

[HKC+09] Stephane Helleringer, Hans-Peter Kohler, Agnes Chimbiri, Praise Chatonda, and James Mkandawire. The likoma network study: Context, data collection, and initial results. *Demographic research*, 21:427, 2009.

[Hoo93] Vera Hoorens. Self-enhancement and superiority biases in social comparison. *European review of social psychology*, 4(1):113–139, 1993.

[ISP+14] Madhura Ingalhalikar, Alex Smith, Drew Parker, Theodore D Satterthwaite, Mark A Elliott, Kosha Ruparel, Hakon Hakonarson, Raquel E Gur, Ruben C Gur, and Ragini Verma. Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828, 2014.

[IVdBV11] Raghuram Iyengar, Christophe Van den Bulte, and Thomas W Valente. Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212, 2011.

[Joh14] Timothy P Johnson. *Handbook of Health Survey Methods*, volume 565. John Wiley & Sons, 2014.

[KBRN14] Ilyana Kuziemko, Ryan W Buell, Taly Reich, and Michael I Norton. âĂIJlast-place aversionâĂİ: Evidence and redistributive implications. *The Quarterly Journal of Economics*, 129(1):105–149, 2014.

[KCM+15] Yury Kryvasheyeu, Haohui Chen, Esteban Moro, Pascal Van Hentenryck, and Manuel Cebrian. Performance of social network sensors during hurricane sandy. *PLoS one*, 10(2):e0117288, 2015.

[KD99] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[KFH+10] Matthias Kowald, Andreas Frei, Jeremy K Hackney, Johannes Illenberger, and Kay W Axhausen. Collecting data on leisure travel: The link between leisure contacts and social interactions. *Procedia-Social and Behavioral Sciences*, 4:38–48, 2010.

[KHS+15] David A Kim, Alison R Hwong, Derek Stafford, D Alex Hughes, A James O'Malley, James H Fowler, and Nicholas A Christakis. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 2015.

[KK07] Geeta Gandhi Kingdon and John Knight. Community, comparisons and subjective well-being in a divided society. *Journal of Economic Behavior & Organization*, 64(1):69–90, 2007.

[KKT03] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[KN11] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[KNT10] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer, 2010.

[Kol09] Eric D Kolaczyk. *Statistical analysis of network data (Springer Series in Statistics)*, volume 69. Springer-Verlag, New York, 2009.

[KPMD+12] Michael W Kraus, Paul K Piff, Rodolfo Mendoza-Denton, Michelle L Rheinschmidt, and Dacher Keltner. Social class, solipsism, and contextualism: how the rich are different from the poor. *Psychological review*, 119(3):546, 2012.

[LAH07] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[Lau73] Edward O Laumann. *Bonds of pluralism: The form and substance of urban social networks*. Wiley-Interscience, 1973.

[LHNB15] Gabriel E Leventhal, Alison L Hill, Martin A Nowak, and Sebastian Bonhoeffer. Evolution and emergence of infectious diseases in theoretical and real-world networks. *Nature communications*, 6, 2015.

[Lin00] Nan Lin. Inequality in social capital. *Contemporary sociology*, 29(6):785–795, 2000.

[LMVdDW11] Jennifer Lindquist, Junling Ma, P Van den Driessche, and Frederick H Willeboordse. Effective degree network disease models. *Journal of mathematical biology*, 62(2):143–164, 2011.

[Lut05] Erzo FP Luttmer. Neighbors as negatives: Relative earnings and well-being. *The Quarterly journal of economics*, 120(3):963–1002, 2005.

[MAC+85] Brian Mullen, Jennifer L Atkins, Debbie S Champion, Cecelia Edwards, Dana Hardy, John E Story, and Mary Vanderklok. The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology*, 21(3):262–283, 1985.

[Mar90] Peter V Marsden. Network data and measurement. *Annual review of sociology*, pages 435–463, 1990.

[Mar03] Peter V Marsden. Interviewer effects in measuring network size using a single name generator. *Social Networks*, 25(1):1–16, 2003.

[Mar11] Peter V Marsden. Survey methods for network data. *The SAGE handbook of social network analysis*, pages 370–388, 2011.

[McC03] Scott D McClurg. Social networks and political participation: The role of social interaction in explaining political participation. *Political research quarterly*, 56(4):449–464, 2003.

[MCF10] Sara C Mednick, Nicholas A Christakis, and James H Fowler. The spread of sleep loss influences drug use in adolescent social networks. *PloS one*, 5(3):e9775, 2010.

[MH07] Alexandra Marin and Keith N Hampton. Simplifying the personal network name generator alternatives to traditional multiple and single name generators. *Field methods*, 19(2):163–193, 2007.

[MKA+01] Mark Montgomery, Gebre-Egziabher Kiros, Dominic Agyeman, John B Casterline, Peter Aglobitse, and Paul C Hewett. *Social networks and contraceptive dynamics in Southern Ghana*. Citeseer, 2001.

[MKC+04] Nilly Madar, Tomer Kalisky, Reuven Cohen, Daniel ben Avraham, and Shlomo Havlin. Immunization and epidemic dynamics in complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):269–276, 2004.

[MNHD+10] Vincent Marceau, Pierre-André Noël, Laurent Hébert-Dufresne, Antoine Allard, and Louis J Dubé. Adaptive networks: Coevolution of disease and topology. *Physical Review E*, 82(3):036116, 2010.

[Mor53] Jacob Levy Moreno. Who shall survive? foundations of sociometry, group psychotherapy and socio-drama. 1953.

[MPSV02] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 26(4):521–529, 2002.

[MRS08] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, UK, 2008.

[MSL87] J Miller McPherson and Lynn Smith-Lovin. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review*, pages 370–379, 1987.

[MSLC01] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[Nar02] Deepa Narayan. Bonds and bridges: social capital and poverty. *Social capital and economic development: well-being in developing countries. Northampton, MA: Edward Elgar*, pages 58–81, 2002.

[New10] Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

[Nir05] Lilach Nir. Ambivalent social networks and their consequences for participation. *International Journal of Public Opinion Research*, 17(4):422–442, 2005.

[NMLB+12] Martial L Ndeffo Mbah, Jingzhou Liu, Chris T Bauch, Yonas I Tekel, Jan Medlock, Lauren Ancel Meyers, and Alison P Galvani. The impact of imitation on vaccination behavior in social contact networks. *PLoS Comput Biol*, 8(4):e1002469, 2012.

[NSRC15] Akihiro Nishi, Hirokazu Shirado, David G Rand, and Nicholas A Christakis. Inequality and visibility of wealth in experimental social networks. *Nature*, 526(7573):426–429, 2015.

[OSH+07] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[PS09] James E Pustejovsky and James P Spillane. Question-order effects in social network name generators. *Social networks*, 31(4):221–229, 2009.

[PSC15] Jessica M Perkins, SV Subramanian, and Nicholas A Christakis. Social networks and health: A systematic review of sociocentric network studies in low-and middle-income countries. *Social Science & Medicine*, 125:60–78, 2015.

[PSCVMV15] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Review of Modern Physics*, 87:925–979, 2015.

[PSS06] Mike Pearson, Christian Sieglich, and Tom Snijders. Homophily and assimilation among sport-active adolescent substance users. *Connections*, 27(1):47–63, 2006.

[PSV01a] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, 63(6):066117, 2001.

[PSV01b] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[PSV02] Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex networks. *Physical Review E*, 65(3):036104, 2002.

[PSV05] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemics and immunization in scale-free networks. *Handbook of graphs and networks: from the genome to the internet*, pages 111–130, 2005.

[Put95] Robert D Putnam. Bowling alone: America's declining social capital. *Journal of democracy*, 6(1):65–78, 1995.

[RAC11] David G Rand, Samuel Arbesman, and Nicholas A Christakis. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108(48):19193–19198, 2011.

[Rap57] Anatol Rapoport. Contribution to the theory of random and biased nets. *The bulletin of mathematical biophysics*, 19(4):257–277, 1957.

[RD02] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70. ACM, 2002.

[Rei99] David A Reingold. Social networks and the employment problem of the urban poor. *Urban Studies*, 36(11):1907–1932, 1999.

[RFC11] J Niels Rosenquist, James H Fowler, and Nicholas A Christakis. Social network determinants of depression. *Molecular psychiatry*, 16(3):273–281, 2011.

[RH61] Anatol Rapoport and William J Horvath. A study of a large sociogram. *Behavioral Science*, 6(4):279–291, 1961.

[RMFC10] J Niels Rosenquist, Joanne Murabito, James H Fowler, and Nicholas A Christakis. The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, 152(7):426–433, 2010.

[RNFC14] David G Rand, Martin A Nowak, James H Fowler, and Nicholas A Christakis. Static network structure can stabilize human cooperation. *Proceedings of the National Academy of Sciences*, 111(48):17093–17098, 2014.

[Rob15] Garry Robins. *Doing Social Network Research: Network-based Research Design for Social Scientists*. Sage, 2015.

[Ros00] Richard Rose. How much does social capital add to individual health? *Social science & medicine*, 51(9):1421–1435, 2000.

[RSE12] Stephanie M Reich, Kaveri Subrahmanyam, and Guadalupe Espinoza. Friending, iming, and hanging out face-to-face: overlap in adolescents' online and offline social networks. *Developmental psychology*, 48(2):356, 2012.

[SBD99] Iris W Schmidt, Ina J Berg, and Betto G Deelman. Illusory superiority in self-reported memory of older adults. *Aging, Neuropsychology, and Cognition*, 6(4):288–301, 1999.

[SCF14] Holly B Shakya, Nicholas A Christakis, and James H Fowler. Association between social network communities and health behavior: an

observational sociocentric network study of latrine ownership in rural india. *American journal of public health*, 104(5):930–937, 2014.

[SDBA12] Alistair Sutcliffe, Robin Dunbar, Jens Binder, and Holly Arrow. Relationships and the social brain: integrating psychological and evolutionary perspectives. *British journal of psychology*, 103(2):149–168, 2012.

[Shu76] Norman Shulman. Network analysis: a new addition to an old bag of tricks. *Acta Sociologica*, 19(4):307–323, 1976.

[Sim55] Georg Simmel. *Conflict and the Web of Group Affiliations*. Free Press, Glencoe, IL, 1922 [1955].

[SLL+14] Jari Saramäki, EA Leicht, Eduardo López, Sam GB Roberts, Felix Reed-Tsochas, and Robin IM Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.

[SMC+13] Michele Starnini, Anna Machens, Ciro Cattuto, Alain Barrat, and Romualdo Pastor-Satorras. Immunization strategies for epidemic processes in time-varying contact networks. *Journal of theoretical biology*, 337:89–100, 2013.

[SR51] Ray Solomonoff and Anatol Rapoport. Connectivity of random nets. *The bulletin of mathematical biophysics*, 13(2):107–117, 1951.

[SS76] FF Strayer and Janet Strayer. An ethological analysis of social agonism and dominance relations among preschool children. *Child Development*, pages 980–989, 1976.

[SSP10] Christian Steglich, Tom AB Snijders, and Michael Pearson. Dynamic networks and behavior: Separating selection from influence. *Sociological methodology*, 40(1):329–393, 2010.

[ST10] Michael Szell and Stefan Thurner. Measuring social dynamics in a massive multiplayer online game. *Social networks*, 32(4):313–329, 2010.

[STRM97] Mark Schneider, Paul Teske, Christine Roch, and Melissa Marschall. Networks to nowhere: Segregation and stratification in networks of information about schools. *American Journal of Political Science*, pages 1201–1223, 1997.

[Taj81] Henri Tajfel. *Human groups and social categories: Studies in social psychology*. Cambridge University Press Archive, 1981.

[TM69] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

[Val03] Thomas W Valente. Social network influences on adolescent substance use: An introduction. *Connections*, 25(2):11–16, 2003.

[Vee91] Ruut Veenhoven. Is happiness relative? *Social indicators research*, 24(1):1–34, 1991.

[VM04] Penny S Visser and Robert R Mirabile. Attitudes in the social context: the impact of social network composition on individual-level attitude strength. *Journal of personality and social psychology*, 87(6):779, 2004.

[VT98] Theo Van Tilburg. Interviewer effects in the measurement of personal network size a nonexperimental study. *Sociological Methods & Research*, 26(3):300–328, 1998.

[WAW+15] Zhen Wang, Michael A Andrews, Zhi-Xi Wu, Lin Wang, and Chris T Bauch. Coupled disease–behavior dynamics on complex networks: A review. *Physics of life reviews*, 15:1–29, 2015.

[WBS+09] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.

[WCW+73] Barry Wellman, Paul Craven, Marilyn Whitaker, H Stevens, A Shorter, S DuToit, and H Bakker. Community ties and support systems: From intimacy to support. *The form of cities in central Canada: Selected papers*, pages 152–167, 1973.

[Wel79] Barry Wellman. The community question: The intimate networks of east yorkers. *American journal of Sociology*, pages 1201–1231, 1979.

[WF94] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[Woo01] Michael Woolcock. The place of social capital in understanding social and economic outcomes. *Canadian journal of policy research*, 2(1):11–17, 2001.

[WS98] Duncan J Watts and Steven H Strogatz. Collective dynamics of âĂŸsmall-worldâĂŹnetworks. *nature*, 393(6684):440–442, 1998.

[ZJ01] Ezra W Zuckerman and John T Jost. What makes you think you're so popular? self-evaluation maintenance and the subjective side of the" friendship paradox". *Social Psychology Quarterly*, pages 207–223, 2001.

[ZSHD05] W-X Zhou, Dider Sornette, Russell A Hill, and Robin IM Dunbar. Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1561):439–444, 2005.