DESCRIPTION, PREDICTION AND EVOLUTION OF A LARGE, DYNAMIC NETWORK FROM INCOMPLETE DATA

A Dissertation Presented

by

Catherine Anne Bliss

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fullfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Mathematical Sciences

May, 2014

Accepted by the Faculty of the Graduate College, The University of Vermont, in partial fulfillment of the requirements for the degree of Doctor of Philosophy, specializing in Mathematical Sciences.

Dissertation Examination Committee:

Co-advisor
Chris Danforth, Ph.D.

Co-advisor
Peter Dodds, Ph.D.

Richard Foote, Ph.D.

Chairperson
Elizabeth Pinel, Ph.D.

Dean, Graduate College
Cynthia J. Forehand, Ph.D.

Date: February 6, 2014

Abstract

Complex networks underlie a variety of social, biological, physical, and virtual systems. Understanding the topology of networks, the manner in which agents interact and evolutionary dynamics of the system can be challenging, both computationally and theoretically. In many settings, network data is incomplete; it is impossible to observe all nodes and all network interactions due to sampling constraints in large datasets or covert interactions between agents.

As both a test of our general methods and as a problem of scientific interest in itself, we focus our attention on over 100 million tweets from the microblogging service Twitter authored between September 2008 and February 2009. This dataset accounts for approximately 30% of all tweets authored in this timespan. The goals of our analysis are threefold: to develop a construction of social networks from replies and reciprocated replies, predict future links in a way that ellucidates evolutionary dynamics, and to scale global statistics of sampled network data to account for incomplete and missing observations.

We begin by defining Twitter reciprocal reply networks and examine the revealed social network structure and dynamics over the time scales of days, weeks, and months. At the level of user behavior, we employ our hedonometric analysis methods to investigate patterns of sentiment expression. We find users average happiness scores to be positively and significantly correlated with those of users one, two, and three links away. We strengthen our analysis by proposing and using a null model to test the effect of network topology on the assortativity of happiness. We also find evidence that more well connected users write happier status updates, with a transition occurring around Dunbar's number. Second, we use an evolutionary algorithm to optimize weights which are used in a linear combination of sixteen neighborhood and node similarity indices to predict future links. Our method exhibits fast convergence and high levels of precision for the top twenty predicted links. Based on our findings, we suggest possible factors which may be driving the evolution of Twitter reciprocal reply networks.

Lastly, we acknowledge that our dataset is incomplete and explore how global network statistics scale with missing data in a variety of sampling regimes. We propose scaling methods to predict true network parameters from only partial knowledge of nodes, links, or weighted interactions. We validate our analytical results with four classes of simulated networks (Erdös-Rényi, Scale-free, Small World, and Range dependent) and six empirical data sets. To overcome limitations due to sampling tweets, we apply our developed methods to Twitter reply networks and suggest a characterization of the Twitter interactome for this time period.

Citations

Material from this dissertation has been published in the following form:

Bliss, C. A., Kloumann, K. I., Harris, K. D., Danforth, C. M., and Dodds, P.S.. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science* **3**(5), 388-397.

AND

Bliss, C. A., Frank, M. R., Danforth, C. M., and Dodds, P.S.. (2014). An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks. *Journal of Computational Science*.

Dedication

in memory of

Dr. DeGraff Everett Bliss, Jr. (1941-1992)

Acknowledgements

I would like to acknowledge the following individuals who made this work possible. First, I wish to thank my dissertation committee chairperson, Elizabeth Pinel, and committee members Chris Danforth, Peter Dodds, and Richard Foote. In particular, I am deeply grateful to my co-advisors, Chris Danforth and Peter Dodds for their inspiration, guidance, and enthusiasm for this work. I also wish to thank Richard Foote for his guidance as my M.S. advisor and continued support and tutelage. Second, I also wish to thank many others, such as Mariarosa Allodi, Hussein Behforooz and Dave Dummit for being pivotal mentors in my Mathematics career and shaping my desire to pursue graduate study in Mathematics. Third, I wish to express gratitude for the Vermont Complex Systems Center, the individuals involved in creating the Center and the CSYS Fellowship which I received from 2009-2011. In particular, I wish to thank Maggie Eppstein, Jim Bagrow and students of the Computational Story Lab for their insightful conversations over the years. Fourth, I wish to acknowledge the Vermont Advanced Computing Core which is supported by NASA (NNX-08AO96G) at the University of Vermont for providing High Performance Computing resources and extend a very special thanks to Jim Lawson for his dedication and helpfulness. Fifth, I wish to acknowledge support from the Mitre Corporation and Graduate Fellowship funding in part from NSF Career Award #0846668 to Peter Dodds. Lastly, I wish to thank my family. To Roger Czupryna, I thank you for reminding me about what matters most in life. To Jose Joaquin Lizano Ortega, I thank you for your encouragement to begin this endeavor and your insistence to enjoy life alongside hard work. To my mother, Sharon Bliss, I recognize that none of this would have been possible without your unwavering support, wisdom, and confidence in my ability to succeed.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Complex networks underlie a variety of social, biological, physical, and virtual systems. Problematically, empirically gathered network data is often incomplete in that not all interactions or entities are observed in sampling. In our work, we develop several tools for describing a large, time-varying social network from only partial knowledge of network interactions. Additionally, we explore predictive tools for network densification and suggest possible mechanisms which may be driving network evolution. Although our efforts largely focus on a particular network of study, our methods are transferable to networks across multiple domains. The remainder of this section introduces our dataset and overviews the aims of our work.

1.1 Twitter as a living laboratory

Twitter is an online, interactive social media platform in which users post tweets, microblogs with a 140 character limit. Since its inception in 2006, Twitter has reached global scale, with over 215 million monthly active users as of October 2013 (Twitter, 2013). Tweets are open online by default, and are also broadcast directly to a user's followers. Users may express interest in a tweet by retweeting the message to their followers. Alternatively, followers may reply directly to the author.

With the abundance of publicly available data and the surge in the number of new accounts, Twitter has come to serve as a living laboratory for studying dynamic social networks (Bakshy et al., 2011; Bollen, Mao & Zeng, 2011; Cha et al., 2010; Dodds et al., 2011; Fischer & Reuber, 2011; Frank et al., 2013; Golder & Yardi, 2010; Gonçalves, Perra, & Vespignanai, 2011; Gruzd, Doiron, & Mai, 2011; Huberman, Romero, & Wu, 2008; Hutto, Yardi & Gilbert, 2013; Java et al., 2009; Kim et al., 2009; Kloumann et al., 2012; Kwak et al., 2010; Mitchell et al., 2013; Morstatter et al., 2013; Romero, Meeder, & Kleinberg, 2011; Romero & Kleinberg, 2010; Romero, Tan, & Ugander, 2013; Rowe, Stankovic, & Alani, 2012; Thelwall, Buckley, & Paltoglou, 2011; Weng et al., 2010).

In our work, we focus our attention on over 100 million tweets from Twitter authored between September 2008 and February 2009. This dataset accounts for varying proportions of all tweets authored in this timespan, as detailed in the subsequent chapters. Our goals in this project are threefold: to describe the construction of social networks from reply messages, predict future links that will occur in a way that elucidates evolutionary dynamics, and to scale global network statistics of sampled network data to account for incomplete data. We now overview each of these goals and note that more detail, including a literature review of related work, is provided in Chapters 2, 3 and 4 respectively.

1.2 Twitter reciprocal reply networks

Social network analysis, a subfield of network science that focuses on social interactions between entities, has a long history in both theoretical and applied settings (Wasseerman & Faust, 1994). Driven by the increased availability of real-time, in-situ data reflecting people's social interactions and choices, there has been an explosion of research activity characterizing large-scale online social networks, such as blogs, Facebook, LinkedIn, and

Twitter (Adamic & Glance, 2005; Bakshy et al. 2011, Bollen, Mao, & Zeng, 2011; Bollen et al., 2011; Cha et al., 2010; Dodds & Danforth, 2010; Dodds et al., 2011; Gjoka et al., 2010; Guo et al., 2009; Huberman, Romero, & Wu, 2008; Java et al., 2009; Kim et al., 2009; Kwak et al., 2010; Papacharissi, 2009; Tan et al., 2011; Thelwall et al., 2011; Ugander et al., 2012; Viswanath et al., 2009; Weng et al., 2010). In this study, we examine the patterns of interactions between individuals using the microblogging service Twitter. Our interest is both as a case of scientific interest, as Twitter has come to serve as peristent and pervasive media, and as a testbed for more general methods which apply more generally to networks in various domains.

The majority of previous studies have examined the topology of and information cascades on the Twitter follower network (Bakshy et al., 2011; Cha et al., 2010; Kwak et al., 2010), as well as on networks derived from mutual following (Bollen, et al., 2011). However, the follower network is not the only representation of Twitter's social network, and its structure can be misleading (Gonçalves, Perra, & Vespignani, 2011). Kwak and others (2010) found very few individuals who followed their followers and questioned the extent to which Twitter exhibits social network characteristics, if at all.

Additional concerns relate to determining how long a link between two users in the network should persist. Including stale user-user interactions in the network mistakenly creates an inaccurate portrayal of the current state of the system; this is typically referred to as the "unfriending problem" (Noel, Galuba, & Nyhan, 2011). Not only will network statistics such as the number of nodes, average degree, maximum degree and proportion of nodes in the giant component be artificially inflated due to superfluous, no-longer-active links (Grannis, 2010; Noel, Galuba, & Nyhan, 2011), but the degree distribution will also be distorted. Kwak et al. (2010) found that the degree distribution for a Twitter follower

network deviated from a power law distribution due to an overabundance of high degree nodes resulting from an accumulation of "dead-weight" in the network.

In Chapter 2, we develop the concept of Twitter reply and reciprocal reply networks and define their construction. Recognizing that, due to practical limitations, accumulation of network data must occur on some scale, we analyze users in day, week, and month reciprocal reply networks. By examining networks constructed from reciprocated communication and at smaller time scales, we aim to take a more dynamic view of the interactions occurring in this network.

Characterizing how ideas and emotions spread through social networks, as well as how individuals self-organize in these settings, can impact society by aiding our understanding of how social media reflects and facilitates social change. Several studies have explored the assortativity of happiness (Bollen et al., 2011; Fowler & Christakis, 2008), obesity (Christakis & Fowler, 2007), smoking (Christakis & Fowler, 2008), alcohol consumption (Rosenquist et al., 2010), and loneliness (Cacioppo, Fowler & Christakis, 2009) in social networks. Some have even gone so far as to assert that these phenomena are contagious (Christakis & Fowler, 2013; Hill et al., 2010), however this work has been critiqued due to the failure to account for homophily and additional complications of incomplete network data (Noel, Galuba, & Nyhan, 2011; Lyons, 2011; Shalizi & Thomas, 2011). The observation that social networks exhibit assortativity with respect to these traits evidently requires further study.

In addition to defining reciprocal reply networks and advocating for their use, we also seek to describe how happiness is distributed in the reciprocal reply networks of Twitter. Previous work by others (Bollen et al., 2011; Christakis & Fowler, 2011) suggests that happiness is assortative in social networks and hedonometric work with Twitter data has

revealed cyclical fluctuations in average happiness at the level of days and weeks, as well as spikes and troughs over a time scale of years corresponding to major holidays, political events and catastrophes (Bollen, Mao & Zeng, 2011; Dodds et al., 2011; Golder & Macy, 2011; Kim et al., 2009; Miller, 2011; Thelwell, Buckley, & Paltoglou, 2011). Chapter 2 describes the application of our recently developed hedonometric analysis (Dodds et al., 2011) to compare individuals' sentiment expression and with that of their neighbors one, two and three links away.

1.3 Link prediction

Time varying social networks track dynamics as they change over time. Individuals, represented by nodes, may enter or exit the network, while interactions, represented by links, may strengthen or weaken. While network growth models focus on global properties, the link prediction problem aims to identify new links which will form in the next timestep, given a snapshot of a network at the current time (Liben-Nowell & Kleinberg, 2007). In recent years, there has been a surge of interest in link prediction, with applications ranging to issues of national security, organizational studies (predicting potential collaborators), and online social networking sites (people you may know). In addition to these goals, the identification of predictors for future link formation may prove fruitful for revealing drivers of network densification and evolution.

Previous link prediction efforts related to Twitter have largely focused on predicting follower relationships (Golder & Yardi, 2010; Hutto, Yardi & Gilbert, 2013; Romero & Kleinberg, 2010; Rowe, Stankovic, & Alani, 2012; Yin, Hong & Davison, 2011). In Chapter 3, we detail these efforts and overview link prediction strategies that have been applied in other domains. As noted by Lu et al. (2010), maximum likelihood algorithms and prob-

abilistic models can be prohibitively time consuming for large networks. Given our interest in large, sparse networks with $N \gtrsim 10^6$, we instead focus primarily on topological-based and node attribute (Table 3.1).

In Chapter 3, we describe several topological and node attribute similarity indices (Adamic & Adar, 2003; Barabási et al. 2002; Katz, 1953; Lichtenwalter, Lussier, & Chawla, 2010; Lin, 1998; Lü & Zhou, 2011; Lu et al., 2010; Newman, 2001b; Ravasz, 2002; Salton & McGill, 1986; Sorensen, 1948; Wang & Rong, 2013; Yang et al., 2012; Zhou, Lü, & Zhang, 2009). Depending on the network under analysis, various measures have shown to be particularly promising (Backstrom & Leskovec, 2011; Esslimani, Brun, & Boyer, 2011; Liben-Nowell & Kleinberg, 2007; Leroy, Cambazoglu, & Bonchi 2010; Wang et al., 2011; Yin, Hong, Davison, 2011; Zhou, Lü, & Zhang, 2009).

Several researchers have used supervised learning algorithms to combine features, such as similarity indices,for link prediction efforts (Backstrom & Leskovec, 2011; Al Hasan, 2006; lichtenwalter, Lussier, & Chawla, 2010; Wang et al., 2011). Of particular interest, Wang et al. (2011) study a network of individuals constructed from mobile phone call data. They compare similarity indices used in isolation to a link predictor combining several indices (binary decision tree determined from supervised learning). These researchers found that the combination of node-specific and topological similarity indices outperform topological indices in isolation. While their results are promising, they acknowledge that the cost comes from looking at only a subset of the large potential set of user-user pairs two-links away.

Motivated by the above, we aim to provide a link predictor encompassing both topological and node-specific information exhibiting fast convergence and which reveals possible mechanisms driving network evolution. To this end, we assume a linear combination of

similarity indices and optimize coefficients using the Covariance Matrix Adaptation Evolution Strategy developed by Hansen and Ostermeier (Hansen & Ostermeier, 2001). Chapter 3 describes the advantages and limitations of this work in greater detail.

1.4 Incomplete network data

In practice, data collected about networks is often incomplete due to covert interactions or constraints in sampling. Particular individuals may wish to remain hidden, such as members of organized crime, and individuals who are otherwise overt may have some interactions that they wish to remain hidden because those interactions are of a sensitive nature (e.g., sexual contacts). In other instances, sampling constraints for extremely large networks necessitate an understanding of how network statistics scale under various sampling regimes (Leskovec & Faloutsos, 2006; Morstatter, 2013). Recognizing that we obtain only a fraction of tweets from Twitter's gardenhose API, we seek to develop scaling methods characterizing Twitter reply networks from this incomplete dataset.

When members of a population are drawn at random, each with equal selection probability, the sample parameter being studied is often a good estimate of the population parameter. Problematically, global statistics of subnetwork data are often not good characterizations of the true network because subsamples can be biased in that some individuals or interactions may be more likely to be selected in a subsample ((Costenbader & Valente, 2003; Frantz, Cataldo, & Carley, 2009; Han et al., 2005; Kossinets, 2006; Lee, Kim, & Jeong, 2006; Stumpf, Wiuf, & May, 2005, Stumpf et al., 2008; Wiuf & Stumpf, 2006). For example, nodes of large degree are more likely to be included in a subnetwork generated by randomly sampled links, as compared to nodes of small degree.

This bias in sampling has been an obstacle, particularly with regard to incomplete network data derived from sampled links or weighted interactions (Kolaczyk, 2009). The obstacle in predicting the number of nodes in a network from only knowledge of a subnetwork generated from sampled links or interactions has been centered around difficulties in predicting the true degree distribution from samplinged network data. Chapter 4 overviews work by others in this area (Frank, 1980, Stumpf et al., 2005), as well as our methods which overcome this obstacle and allow us to characterize the Twitter interactome.

Our work concludes with a description of Twitter reply networks. We find evidence of an upper limit for the number of links an individual can actively engage in communication, providing further support for Dunbar's hypothesis in Twitter reply networks (Dunbar, 1995). We hypothesize that Twitter reply networks evolve with constraints whereby new links form in accordance with limits to time and attention (e.g. Resource Allocation).

Chapter 2

Twitter reciprocal reply networks exhibit assortativity with respect to happiness

The advent of social media has provided an extraordinary, if imperfect, big data window into the form and evolution of social networks. Based on nearly 40 million message pairs posted to Twitter between September 2008 and February 2009, we construct and examine the revealed social network structure and dynamics over the time scales of days, weeks, and months. At the level of user behavior, we employ our recently developed hedonometric analysis methods to investigate patterns of sentiment expression. We find users average happiness scores to be positively and significantly correlated with those of users one, two, and three links away. We strengthen our analysis by proposing and using a null model to test the effect of network topology on the assortativity of happiness. We also find evidence that more well connected users write happier status updates, with a transition occurring around Dunbar's number. More generally, our work provides evidence of a social sub-network structure within Twitter and raises several methodological points of interest with regard to social network reconstructions.

2.1 Introduction

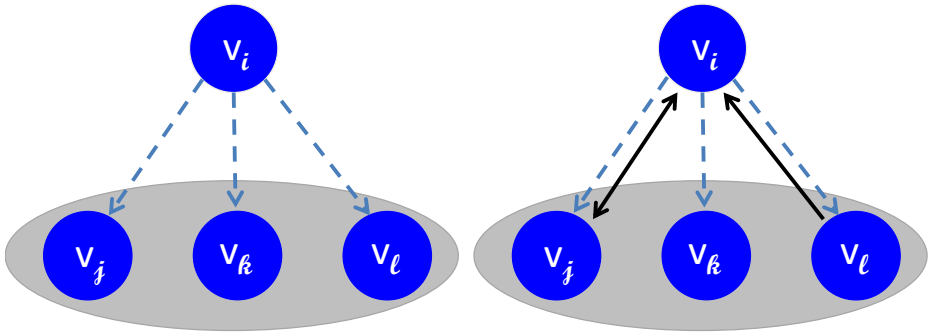
Social network analysis has a long history in both theoretical and applied settings [1]. During the last 15 years, and driven by the increased availability of real-time, in-situ data reflecting people's social interactions and choices, there has been an explosion of research activity around social phenomena, and many new techniques for characterizing large-scale social networks have emerged. Numerous studies have examined the structure of online social networks in particular, such as blogs, Facebook, and Twitter [2–19].

In a series of analyses of the Framingham Heart Study data and the National Longitudinal Study of Adolescent Health, Christakis, Fowler, and others have examined how qualities such as happiness, obesity, disease, and habits (e.g., smoking) are correlated within social network neighborhoods [20–25]. The authors' additional assertion of contagion, however, has been criticized primarily on the basis of the difficulties to be found in distinguishing these phenomena from homophily [26–28]. The observation that social networks exhibit assortativity with respect to these traits evidently requires further study and leads us to explore potential mechanisms. Advances would naturally provide further insight into the nature of how social groups influence individual behavior and vice versa.

Our focus in the present work is the social network of Twitter users. With the abundance of available data, Twitter serves as a living laboratory for studying contagion and homophily [29]. As a requisite step towards these goals, we first define sub-networks of Twitter users suitable to such study and, second, examine whether assortativity is observed in these sub-networks. Before describing our methods, we provide a brief overview of Twitter, related work, and the challenges associated with social network analysis in this arena.

Twitter is an online, interactive social media platform in which users post tweets, microblogs with a 140 character limit. Since its inception in 2006, Twitter has grown to encompass over 200 million accounts, with over 100 million of these accounts currently active as of October 2011, and with some users having garnered over 10 million followers [30]. Tweets are open online by default, and are also broadcast directly to a user's followers. Users may express interest in a tweet by retweeting the message to their followers. Alternatively, followers may reply directly to the author.

Understanding the topology of the Twitter network, the manner in which users interact and the diffusion of information through this media is challenging, both computationally and theoretically. One of the central issues in characterizing the topology of any network representation of Twitter lies in defining the criteria for establishing a link between two users. The majority of previous studies have examined the topology of and information cascades on the Twitter follower network [7, 10, 15], as well as on networks derived from mutual following [8]. However, the follower network is not the only representation of Twitter's social network, and its structure can be misleading [31]. For example, in a study of over 6 million users, Cha et al. [10] found that users with the highest follower counts were not the users whose messages were most frequently retweeted. This suggests that such popular users (as measured by follower count) may not be the most influential in terms of spreading information, and this calls into question the extent to which users are influenced by those that they follow [32]. Of further concern is the finding of low reciprocity within follower networks. Kwak et al. found very few individuals who followed their followers [15]. As a result, trying to infer meaningful influence and contagion in such a network is difficult.



(a) Followers (b) Interaction

Figure 2.1: Depiction of follower networks and reciprocal reply networks. (a) Follower network: The follower network is generated by declared following choices, absent any messages being sent. If user v_i broadcasts tweets to followers v_j, v_k and v_ℓ (represented by the dashed, blue arrow) v_i would be connected to each of v_j, v_k and v_ℓ by a directed link in a follower network. (b) Reciprocal-reply network: Directed replies are represented by a solid black arrow. When considering the interaction between users, a reply (i.e., v_ℓ replies to v_i) provides evidence of a directional interaction between nodes. We mandate a stronger condition for interaction, namely reciprocal replies (i.e., v_j replies to v_i and vice versa) over a given time period. Thus v_i and v_j are connected in the reciprocal reply network that we construct.

While popular users and their many followers clearly exhibit an affiliation, they do not necessarily interact, as there are different relationships implicated by broadcasting (tweeting), sending a message (@someone), and replying to a message. As an example, we consider a user represented by node v_i which has three followers, represented by v_j, v_k, and v_ℓ as shown in Fig. 2.1a. When a user broadcasts tweets to their many followers, as represented by the directed arrow in Fig. 2.1a, this does not imply that followers read or respond to these tweets. Followers v_j, v_k, and v_ℓ receive all tweets broadcast by node v_i, but this provides no guarantee of interaction. Suppose, though, that we observe that v_ℓ replies to v_i as shown in Figure 2.1b. This provides evidence (but not proof) that the user represented by v_ℓ has indeed received a tweet from v_i and is sufficiently motivated to create a response to v_i. Although a directional network based on these replies can be created, such a directional interaction, however, does not suggest reciprocity between the nodes. In this example, we have no evidence that v_i has, in any way, considered or even read such a response from his/her follower.

We conclude that following and unreciprocated replies are not sufficient for interaction and present an alternative means by which to derive a social network from Twitter messages, via reciprocal replies. In our reciprocal-reply network, two nodes, v_i and v_j, are connected if v_i has replied to v_j and v_j has replied to v_i at least once within a given time period of consideration. In Figure 2.1b, the nodes v_i and v_j meet this criterion.

Another challenge in characterizing the topology of any network representation of Twitter concerns determining how long a link between two users in the network should persist. Including stale user-user interactions in the network mistakenly creates an inaccurate portrayal of the current state of the system; this is typically referred to as the "unfriending problem" [26]. Not only will network statistics such as the number of nodes, average

degree, maximum degree and proportion of nodes in the giant component be artificially inflated due to superfluous, no-longer-active links [26, 33], but the degree distribution will also be distorted. Kwak et al. [15] found that the degree distribution for a Twitter follower network deviated from a power law distribution due to an overabundance of high degree nodes resulting from an accumulation of "dead-weight" in the network.

Additional problems are encountered if one uses accumulated network data to measure assortativity with respect to a trait (e.g., happiness). As an example, consider a network in which two users are connected because they interacted during the last week of a year-long study. Including this user-user pair in the list of pairs to compute assortativity for the entire network blurs the relationship between more consistent and repeated interactions that occurred throughout the timespan of the study. Further complications arise when averaging a user's trait over a large time scale (i.e., averaging happiness over a 6 month or 12 month timespan). Detecting changes in users' traits over time and how these may (or may not) be correlated with nearest neighbors' traits is of fundamental importance; accumulated network data occludes exactly the interactions we are looking to understand. Recognizing that, due to practical limitations, accumulation of network data must occur on some scale, we analyze users in day, week, and month reciprocal reply networks. By examining networks constructed at smaller time scales and calculating users' happiness scores based on tweets made only during that time period, we aim to take a more dynamic view of the network.

In addition to defining reciprocal reply networks and advocating for their use, we also seek to describe how happiness is distributed in the reciprocal reply networks of Twitter. Previous hedonometric work with Twitter data has revealed cyclical fluctuations in average happiness at the level of days and weeks, as well as spikes and troughs over a time scale of years corresponding to events such as U.S. Presidential Elections, the Japanese tsunami

and major holidays [11, 34, 35]. Other studies have examined changes in valence of tweets associated with the death of Michael Jackson [14], changes in the U.S. Stock Market [9], the Chilean Earthquake of 2010, and the Oscars [16]. In the present work, we seek to understand localized patterns of happiness in the Twitter users' social network.

Understanding how emotions are distributed through social networks, as well as how they may spread, provides insight into the role of the social environment on individual emotional states of being, a fundamental characteristic of any sociotechnical system. Bollen et al. [8] examine a reciprocal-follower network using Twitter and suggest that Subjective Well-Being (SWB), a proxy for happiness, is assortative. Building on their work, we address whether happiness is assortative in reciprocal-reply networks. We also test the hypothesis of Christakis and Fowler [25] who find evidence that the assortativity of happiness may be detected up to three links away. In doing so, we raise an additional point which is not specific to Twitter networks, but rather relates to empirical measures of assortativity in general. Relatively few studies have employed a null model for calculating the pairwise correlations (e.g., happiness-happiness). We devise a null model which maintains the topology of the network and randomly permutes happiness scores attached to each node. By randomly permuting users' happiness scores, we can detect what effect, if any, network structure has on the pairwise correlation coefficient.

We organize our paper as follows: In Section 2, we describe our data set, the algorithm for constructing reciprocal-reply networks, network statistics used for characterizing the networks, and our measure for happiness. We propose an alternative means by which to detect social structure and argue that our method detects a large social sub-network on Twitter. In Section 3, we describe the structure of this network, the extent to which it is

assortative with respect to happiness and the results of testing assortativity against a null model. In Section 4, we discuss these findings and propose further investigations of interest.

2.2 Methods

2.2.1 Data

From September 2008 to February 2009, we retrieved over 100 million tweets from the Twitter streaming API service.[1] While the volume of our feed from the Twitter API increased during this study period, the total number of tweets grew at a faster rate (Fig. 2.2). During this time period, we estimate that we collected roughly 38% of all tweets.[2] The number of messages and percent of which were replies are reported in Table 2.A4. For the remainder of this paper, we restrict our attention to the nearly 40 million message-reply pairs within this data set and the users who authored these tweets.

The data received from the Twitter API service for each tweet contained separate fields for the identification number of the message (message id), the identification number of the user who authored the tweet (user id), the 140 character tweet, and several other geo-spatial and user-specific metadata. If the tweet was made using Twitter's built-in reply function,[3] the identification number of the message being replied to (original message id) and the identification of the user being replied to (original user id) were also reported.

[1] Data was received in XML format.

[2] We calculated the total number of messages as the difference between the last message id and the first message id that we observe for a given week. This provides a reasonable estimate of the number of tweets made per week, as message ids were assigned (by Twitter) sequentially during the time period of this study.

[3] Twitter has a built-in reply function with which users reply to specific messages from other users. Tweets constructed using Twitter's reply function begin with '@username', where 'username' is the Twitter handle of the user being replied to; the user and message ids of the tweet being replied to are included in the reply message's metadata from the Twitter API. Users often informally reply to or direct messages to other users by including said users' Twitter handles in their tweets. In such cases, however, no identification information about the "mentioned" user is included in the API parameters for these tweets (only their Twitter handle is) and we exclude such exchanges when building the reciprocal reply network.

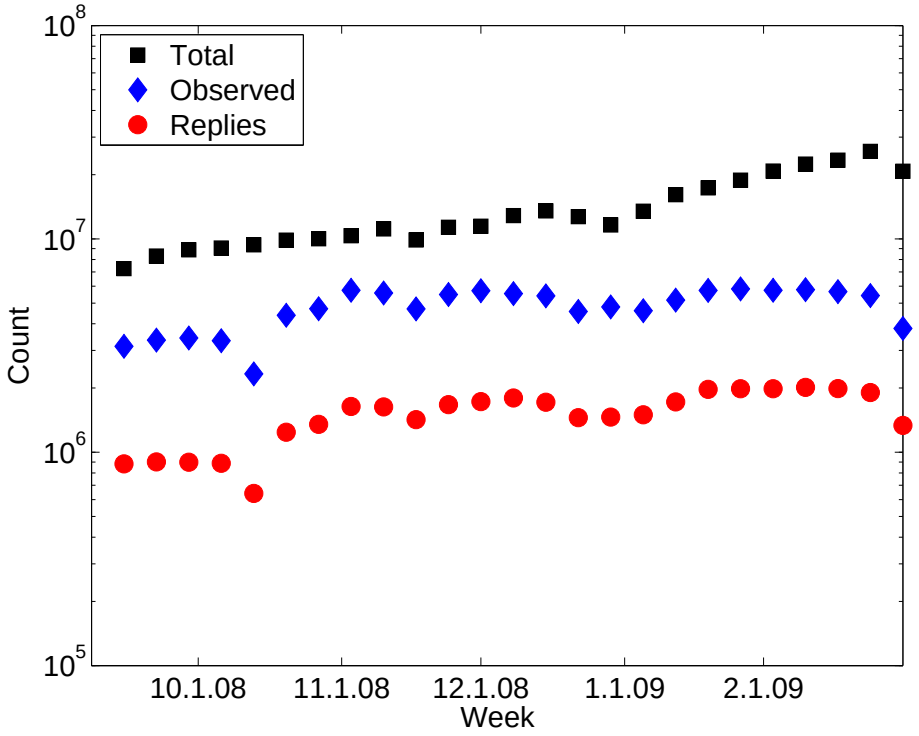
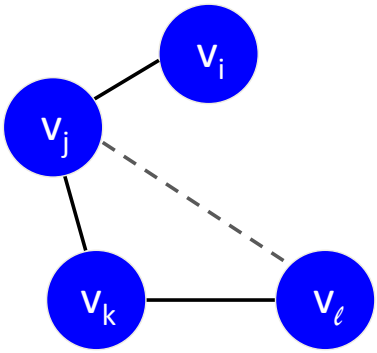


Figure 2.2: Tweet counts for the weeks between September 2008 and February 2009. The three curves represent the total, those that we observed and the number of the observed tweets that constituted replies.

We acknowledge two sources of missing data. First, the Twitter API did not allow us access to all tweets posted during the 6 month period under consideration. Thus, there are replies that we have not observed. As a result, some users may remain unconnected or connected by a path of longer length due to missing intermediary links in our reciprocal-reply network (Fig. 2.3). Secondly, we acknowledge that users may be interacting with each other and not using the built-in reply function. We discuss this further in the next section.



	1-link pairs	2-link pairs	3-link pairs
Observed	(v_i,v_j) (v_j,v_k) (v_k,v_ℓ)	(v_i,v_j) (v_j,v_ℓ)	(v_i,v_ℓ)
True	(v_i,v_j) (v_j,v_k) (v_k,v_ℓ) (v_j,v_ℓ)	(v_i,v_j) (v_i,v_ℓ)	None

Figure 2.3: Effect of missing links in the reciprocal reply network. The effect of missing data in the reciprocal reply network is depicted where observed links are shown as a solid line and an unobserved link is shown as a dashed line. The effect of unobserved links is twofold: (1) some connections between nodes are missed (e.g., v_j and v_ℓ are not connected in the observed reciprocal reply network); and (2) some path lengths between nodes are artificially inflated (e.g., the distance from v_i to v_ℓ is 3 in the observed reciprocal-reply network, however in reality the path length is 2).

2.2.2 Reciprocal-reply network

In keeping with terminology used in the field of complex networks, the terms *nodes* and *links* will be used henceforth to describe users and their connections. Define $\mathcal{G} = (V, E)$ to be a simple graph which contains, $N = |V|$ nodes and $M = |E|$ links. We construct

the reciprocal-reply networks in which users are represented by nodes, $v_i \in V$, and links connecting two nodes, $e_{ij} \in E$, indicate that v_i and v_j have made replies to each other during the period of time under analysis (Fig. 2.1). For each network, we remove self-loops (i.e., users who responded to themselves). We characterize the reciprocal-reply network for each week by the calculation of network statistics such as N (the number of nodes), $\langle k \rangle$ (average degree), $k_{\max}$ (maximum degree), the number of connected components and S (proportion of nodes in the giant component). We calculate clustering, C_G, according to Newman's global clustering coefficient [36]:

$$C_G = \frac{3 \times (\text{number of triangles on a graph})}{\text{number of connected triples of nodes}}.$$

Assortativity refers to the extent to which similar nodes are connected in a network. Often, degree assortativity is quantified by computing the Pearson correlation coefficient of the degrees at each end of links in the network [37]. Since we are interested in quantifying the extent to which the highest degree nodes are connected to other high degree nodes, as defined by the rank of their degrees, we instead measure degree assortativity by the Spearman correlation coefficient.[4] Thus for each link that connects nodes v_i and v_j, we examine the ranks of k_{v_i} and k_{v_j}. The Spearman correlation coefficient, which is the Pearson correlation coefficient applied to the ranks of the degrees at each end of links in the network, is a non-parametric test that does not rely on normally distributed data and is much less sensitive to outliers.[5]

[4] We present both the Spearman and Pearson correlation coefficient in the Appendix, Figure A2. Pearson's correlation coefficient is more sensitive to extreme values and thus obscures the trend in the data, namely that the network is assortative with respect to the rank (i.e., ordering) of nodes' degrees.

[5] Our degree distribution is not Gaussian, as can be seen from Figure 2.7.

In addition, we also investigate user pairs which are connected by a minimal path length of two (or three) in the reciprocal reply networks. We define $d(v_i, v_j)$ to be the path length (i.e., number of links) between nodes v_i and v_j such that no shorter path exists. As a consequence of missing messages, we recognize that some users will appear to remain unconnected or connected by a path of longer length. Figure 2.3 depicts the effect of missing links on inferred path lengths between nodes in the network. Nodes v_j and v_ℓ are adjacent in the network, however, due to the missing link represented by the dashed line, these nodes are inferred to be two links apart.

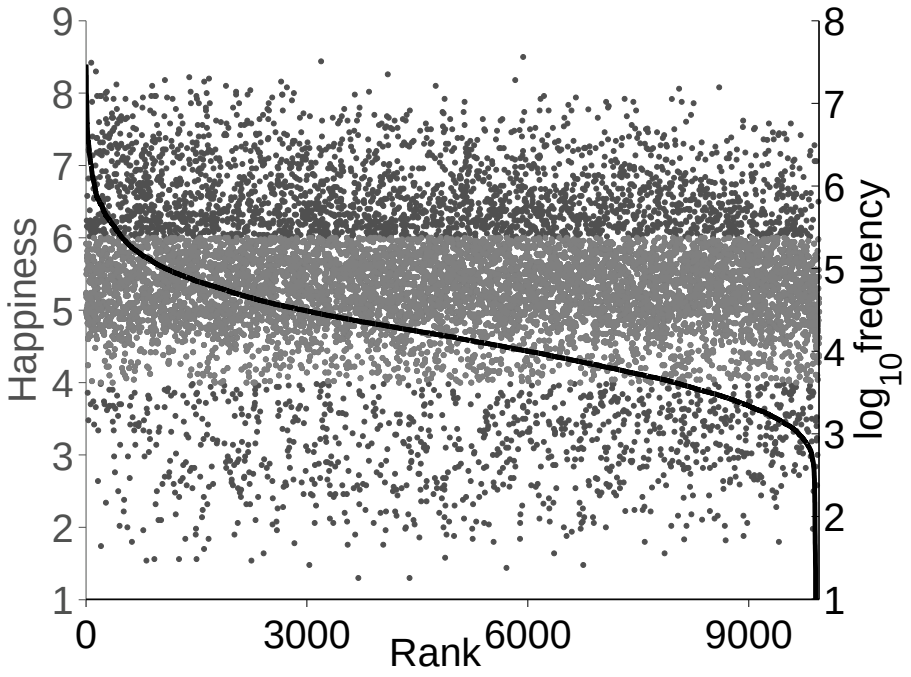


Figure 2.4: The happiness scores of words plotted as a function of their rank. The stop words (words within $\pm\Delta h = 1$ of $h_{\text{avg}} = 5$) are depicted in light grey [38]. These words were excluded from the happiness score computation. The frequency of words and their rank (1=most frequent, 9956=least frequent) are plotted (solid curve). Not all 10,222 labMT words were observed during the time period from September 2008-February 2009.

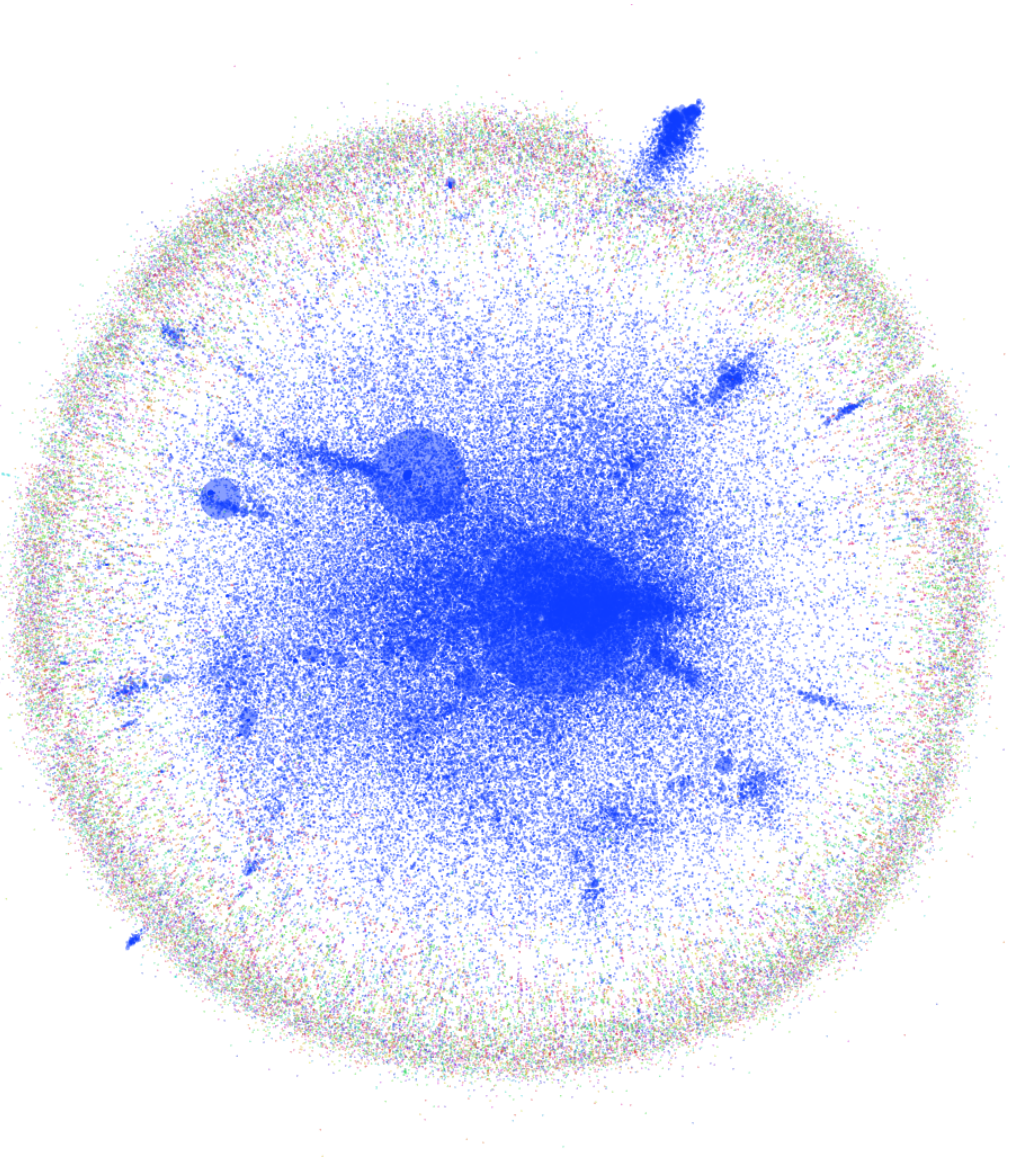


Figure 2.5: Visualization of the components of a reciprocal reply network for one week. A visualization of the 162,445 nodes in the reciprocal reply network for the week beginning December 9, 2008 (Week 14) is shown. Node colors represent connected components, a total of 15342, with the giant component (shown in blue) comprising 76 % of all nodes. The size of each node is proportional to its degree. The visualization was made using Gephi [39].

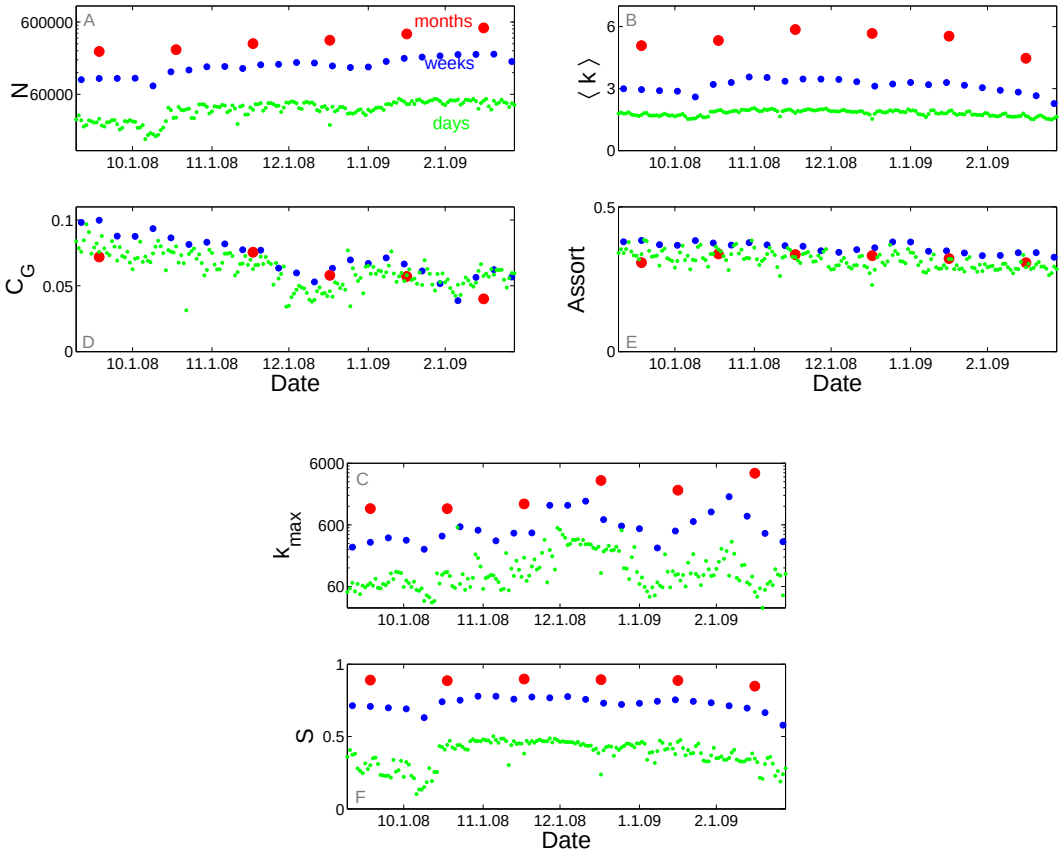


Figure 2.6: Network statistics for the reciprocal-reply networks. (A.) The number of users (N) engaged in reciprocal exchanges when viewed at the level of days (green), weeks (blue), and months (red) increases over the study period. (B.) The average degree ($\langle k \rangle$) remains fairly constant throughout the study period, with higher values detected for larger interaction time periods. (C.) The maximum degree ($k_{\max}$) shows variability throughout the study period. (D.) Clustering decreases quite likely resulting from the inability of the networks' closed triangles to keep up with the growing number of nodes. (E.) Degree assortativity remains fairly constant throughout the study period, and shows little sensitivity to the time period over which the networks represent interactions. (E.) The proportion of nodes in the giant component (S) remains fairly constant for week and month networks, however, shows some variability during the first month of the study for day networks.

2.2.3 Measuring happiness

To quantify happiness for Twitter users, we apply the real-time hedonometer methodology for measuring sentiment in large-scale text developed in Dodds et al. [11]. In this study, the 5000 most frequently used words from Twitter, Google Books (English), music lyrics (1960 to 2007) and the New York Times (1987 to 2007) were compiled and merged into one list of 10,222 unique words.[6] This word list was chosen solely on the basis of frequency of usage and is independent of any other presupposed significance of individual words. Human subjects scored these 10,222 words on an integer scale from 1 to 9 (1 representing sad and 9 representing happy) using Mechanical Turk. We compute the average happiness score (h_{avg}) to be the average score from 50 independent evaluations. Examples of such words and their happiness scores are: h_{avg}(love)=8.42, h_{avg}(special)=7.20, h_{avg}(house)=6.34, h_{avg}(work)=5.24, h_{avg}(sigh)=4.16, h_{avg}(never)=3.34, h_{avg}(sad)=2.38, h_{avg}(die)=1.74. Words that lie within $\pm\Delta h_{\text{avg}} = 1$ of h_{avg}=5 were defined as "stop words" and excluded to sharpen the hedonometer's resolution.[7] The result is a list of 3,686 words, hereafter referred to as the Language Assessment by Mechanical Turk (labMT) word list [11]. See Tables A1 and A2 for additional example word happiness scores.

Figure 2.4 presents word happiness as a function of usage rank for the roughly 10,000 words in the labMT data set. This figure reveals a frequency independent bias towards the usage of positive words (see [37] for further discussion of this positivity bias). Proceeding with the labMT word list, a pattern-matching script evaluated each tweet for the frequency

[6] We provide a brief summary of this methodology here and refer the interested reader to the original paper for a full discussion. The supplementary information contains the full word list, along with happiness averages and standard deviations for these words [11].

[7] For notational convenience, we henceforth use Δh in lieu of Δh_{avg}.

w_i	$h_{\text{avg}}(w_i)$	labMT?	f_i	p_i
Vacation	7.92	yes	1	$\frac{1}{2}$
starts	5.96	yes	n/a	n/a
today	6.22	yes	1	$\frac{1}{2}$
yeahhhhh	n/a	no	n/a	n/a

Table 2.1: Computation of happiness scores. Happiness scores are computed as a weighted average of words' h_{avg} scores. Since "starts" is a stop word, it is not included in the calculation of $h_{\text{avg}}(T) = 7.07$. This example serves is included as a means to illustrate the methodology; in practice, the average is calculated over a much larger word set.

of words. We compute the happiness of each user by applying the hedonometer to the collection of words from all tweets authored by the user during the given time period. Note that each users' collection of words likely reflects messages that were not replies. The happiness of this collection of words is taken to be the frequency weighted average of happiness scores for each labMT word as $h_{\text{avg}}(T) = \frac{\sum_{i=1}^{N} h_{\text{avg}}(w_i) f_i}{\sum_{i=1}^{N} f_i} = \sum_{i=1}^{N} h_{\text{avg}}(w_i) p_i$, where $h_{\text{avg}}(w_i)$ is the average happiness of the ith word appearing with frequency f_i and where p_i is the normalized frequency ($p_i = \frac{f_i}{\sum_{j=1}^{N} f_j}$). As a simple example example, we consider the phrase: *Vacation starts today, yeahhhhh!* in Table 1. In practice, though, the hedonometer is applied to a much larger word set and is not applied to single sentences.

Having found happiness scores for each node (user), we then form happiness-happiness pairs (h_{v_i}, h_{v_j}), where h_{v_i} and h_{v_j} denote the happiness of nodes v_i and v_j connected by an edge. The Spearman correlation coefficient of these happiness-happiness pairs measures how similar individuals' average happiness is to that of their nearest neighbors'. Lastly, we investigate the strength of the correlation between users' average happiness scores and those of other users in the two and three link neighborhoods.

2.3 Results

2.3.1 Reciprocal-reply network statistics

Visualizations of day and week networks were created using the software package Gephi [39]. Figures 2.5 and 2.A6 show a sample week and day network, respectively. All layouts were produced using the Force Atlas 2 algorithm, which is a spring based algorithm that plots nodes together if they are highly connected (see [40] for more details). The sizes of the nodes are proportional to the degrees.

Network statistics, such as the number of nodes (*N*), the average degree $\langle k \rangle$, the maximum degree ($k_{\max}$), global clustering C_G, degree assortativity ($Assort$), and the proportion of nodes in the giant component (S) are summarized in Figure 2.6. Several trends are apparent.

Throughout the course of the study, the number of users in the observed reciprocal-reply network shows an increase, whereas the average degree, degree assortativity, and proportion of nodes in the giant component remain fairly constant. The fluctuations in maximum degree are the result of celebrities or companies having bursts of high volume reply exchanges with their fans during a particular week, for example Bob Bryar, Drummer for the band *My Chemical Romance* ($k_{\max} = 1244$, Week 12), *Namecheap* domain registration company ($k_{\max} = 1245$, Week 13), Twitter's own *Shorty Awards* ($k_{\max} = 1456$, Week 14), and Stephen Fry, actor and mega-blogger ($k_{\max} = 1718$, Week 22). This observation highlights the importance of examining network data on the appropriate time scale, otherwise information about these kinds of dynamics would be be lost. The clustering coefficient shows a slight decrease over the course of this period. This is most likely due to an in-

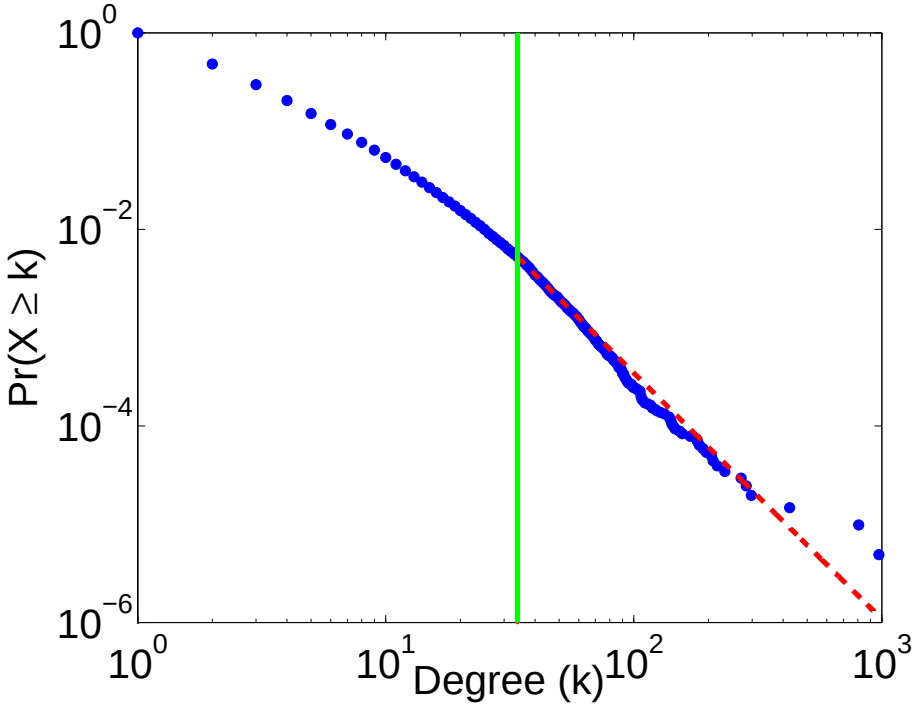


Figure 2.7: Degree distribution for a sample week. Log-log plot of the complementary cumulative distribution function (CCDF) of the degree distribution for a sample week (week of January 27, 2009) network is shown (blue), along with the best fitting power law model ($\alpha = 3.50$ and $k_{\min} = 34$) using the procedure of Clauset, Shalizi, and Newman [41]. We test whether the empirical distribution is distinguishable from a power law using the Kolmogorov-Smirnov test and find no evidence against the null hypothesis ($D = 2.28 \times 10^{-2}, p = 0.095, n = 203852$).

creasing number of nodes, which results in a smaller proportion of closed triangles in the network.

The degree distribution, P_k, for a sample week (week beginning January 27, 2009) is presented in Figure 2.7. Using the approach outlined by Clauset, Shalizi, and Newman [41], we find a lower bound for the scaling region to be $k_{\min} \approx 34$ and a very steep scaling exponent of $\alpha = 3.5$. This suggests a constrained variance and mean. We test whether the empirical distribution is distinguishable from a power law using the Kolmogorov-Smirnov test and find no evidence against the null hypothesis for the week ($D = 2.28 \times 10^{-2}, p = 0.095, n = 203852$). We find the same exponent and statistically stronger evidence of a power law for a sample month (see the Appendix, Fig. 2.A1). This suggests that these distributions' tails may be fit by a power law.

2.3.2 Measuring happiness

The application of the hedonometer gives reasonable results when applied to a large body of text, but can be misleading when applied to smaller units of language [11]. To provide a sense of how sensitive this measure is to the number of labMT words posted by users, we sampled happiness-happiness pairs, (h_{v_i}, h_{v_j}) whose respective users, v_i and v_j, had posted at least α total labMT words during a sample week (week beginning January 27, 2009). For these users, we compute happiness assortativity and show the variation with α in Figure 2.8. For $\Delta h = 0$, there is less variation due to the numerous words centered around the mean happiness score regardless of the threshold, α. Tuning both parameters too high results in few sampled words and corrupts the interpretation of the results.

Figures 2.9 and 2.10 reveal a weakening happiness-happiness correlation for users in the week networks as the path length between nodes increases. All correlations, for each

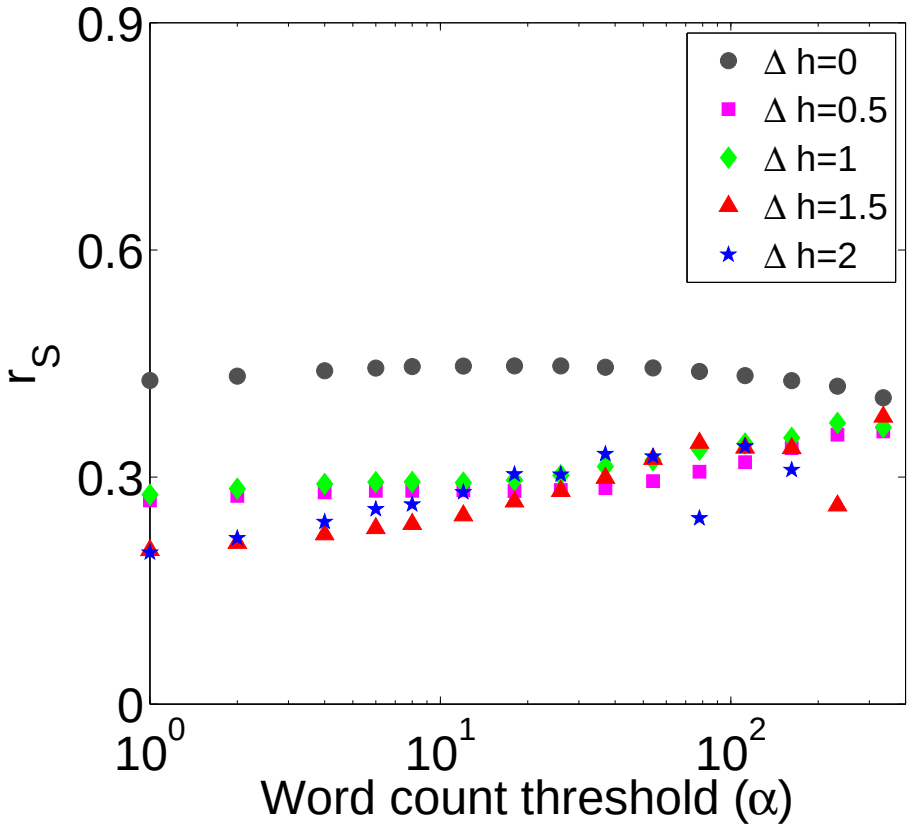


Figure 2.8: Nearest neighbor happiness assortativity. Happiness assortativity as a function of the number of labMT words required per user is displayed for a sample week reciprocal-reply network. Notice that when $\Delta h = 0$, there is less variation due to the numerous words centered around the mean happiness score regardless of the threshold, α. While this stability is desirable, tuning Δh allows us to sharpen the resolution of the hedonometer. This tuning, however, must be balanced with the appropriate choice of α.

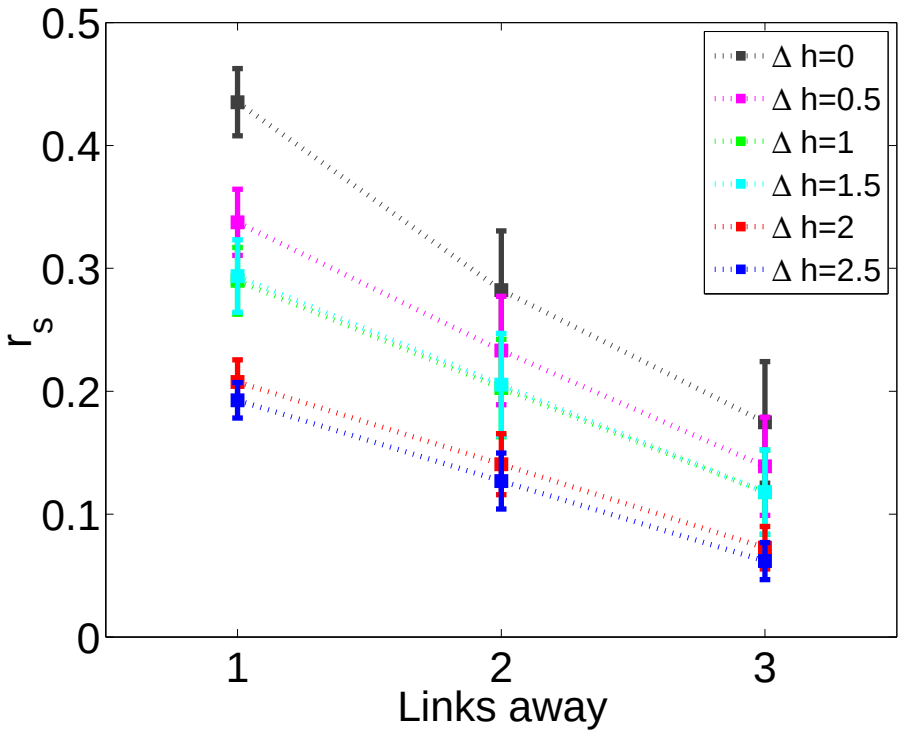
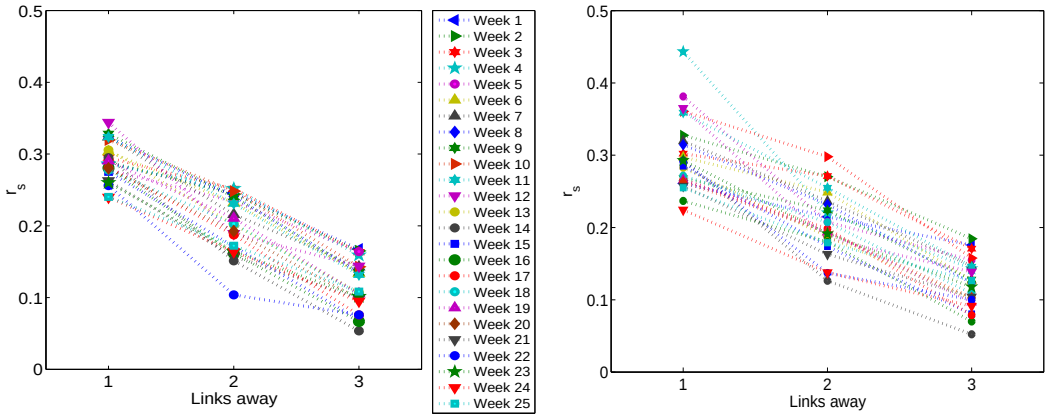


Figure 2.9: Average assortativity of happiness Δh. Average assortativity of happiness for week networks measured by Spearman's correlation coefficients as Δh is dialed from 0 to 2.5, with $\alpha = 50$. As Δh increases, the average correlation decreases. For large Δh the resulting words under analysis have more disparate happiness scores and thus the correlations between users' happiness scores are smaller. Similarly, choosing Δh to be too small (e.g., $\Delta h = 0$) could result in an over estimate of happiness-happiness correlations because of the uni-modal distribution of h_{avg} for the labMT words. Thus a moderate value for Δh is chosen (Δh is set to 1 for this study).



(a) $\Delta h = 1, \alpha = 1$ (b) $\Delta h = 1, \alpha = 50$

Figure 2.10: Happiness assortativity with varying path length. The assortativity of happiness as measured by Spearman's correlation coefficients is shown for week networks, with $\Delta h = 1$ and (a) the threshold of labMT words written by users set to $\alpha = 1$ and (b) $\alpha = 50$. The dashed lines indicate weakening happiness-happiness correlations as the path length increases from one, two, and three links away, for each week in the data set.

week, were significant ($p < 10^{-10}$). This suggests that the network is assortative with respect to happiness and that user happiness is more similar to their nearest neighbors than those who are 2 or 3 links away.

In Figure 2.11 we provide a visualization of an ego-network for a single node, including neighbors up to three links away. Nodes are colored by their h_{avg} score, illustrating the assortativity of happiness. Figure 2.A5 visualizes the happiness assortativity for an entire week network.

In Figure 2.12, we show the average happiness score as a function of user degree k for all week networks. The average happiness score increases gradually as a function of degree, with large degree nodes demonstrating a larger average happiness than small degree nodes. Large degree nodes use words such as "you," "thanks," and "lol" more frequently than small degree nodes, while the latter group uses words such as "damn," "hate," and "tired" more frequently. A word shift diagram, comparing nodes with $k < 100$ vs. nodes

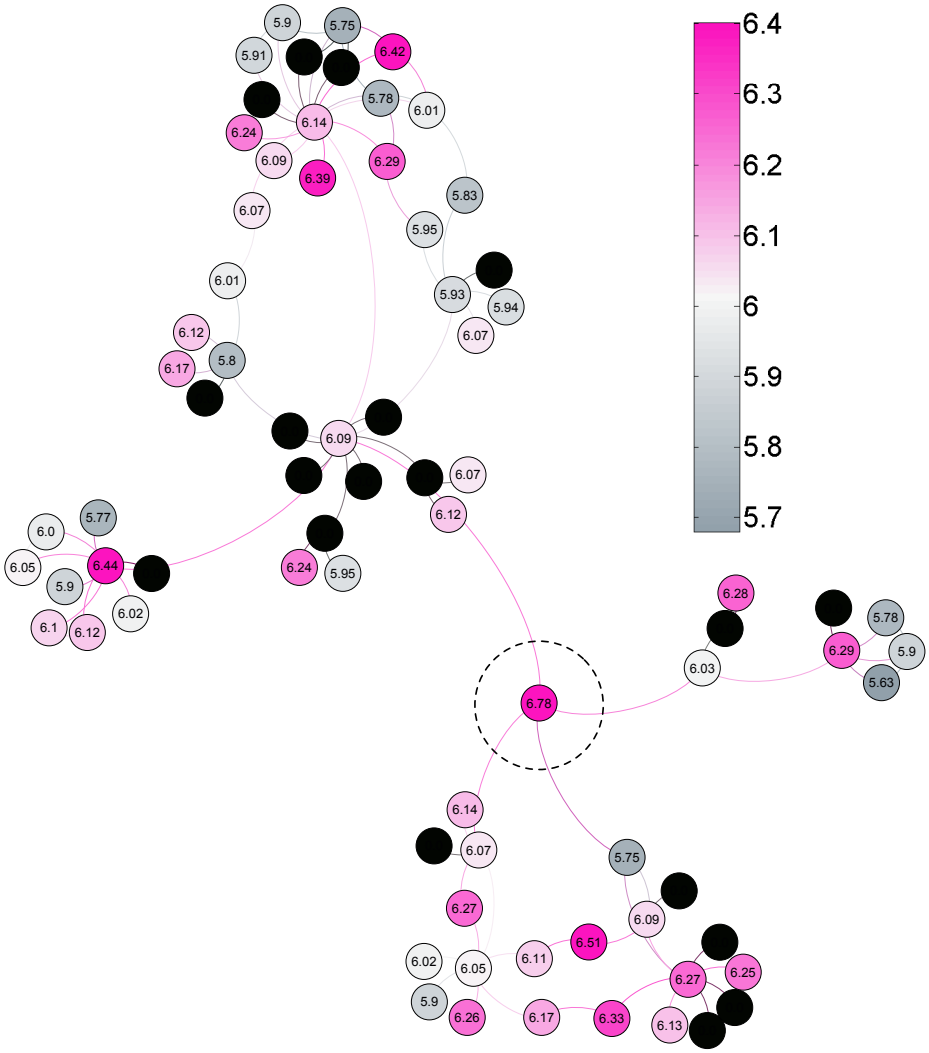


Figure 2.11: Ego-network happiness scores visualization. A visualization of a user and its neighbors 3-links away for a week beginning September 9, 2008 (Week 1). Colors represent happiness scores for users posting more than $\alpha = 50$ labMT words. Nodes depicted with the color black are nodes for which the user's wordbag did not meet our thresholding criteria.

with $k \geq 100$ is included in the Appendix (Fig. 2.A7). Figure 2.12 also reveals that the number of large degree nodes is fairly small. Our results support recent work showing that most users of Twitter exhibit an upper limit on the number of active interactions in which they can be engaged [31]. This may provide further evidence in support of Dunbar's hypothesis, which suggests that the number of meaningful interactions one can have is near 150 [42].

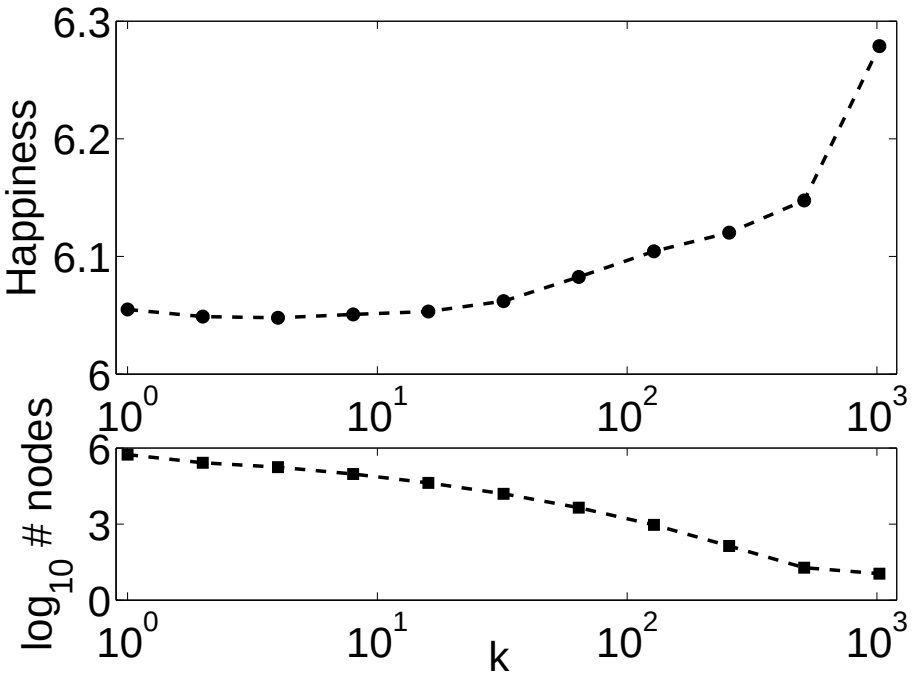


Figure 2.12: Node degree vs. happiness. Top Panel: The average happiness score as a function of user degree k for week networks is increasing, as larger degree nodes use fewer negative words (see Figure 2.A7). Bottom Panel: The number of unique users is reported with respect to degree k; some users appear in more than one bin because they exhibit different degree k for different weeks of the study.

2.3.3 Testing assortativity against a null model

To further examine these findings, we create a null model which maintains the network topology (i.e., adjacency matrices for one link, two link, and three link remain intact), but randomly permutes the happiness scores associated with each node. The Spearman correlation coefficient shows no statistically significant relationship for the null model applied to a sample week of the data set. Figure 2.13 shows the results of 100 random permutations applied to nodes' associated happiness scores. The Spearman correlation coefficients for the observed data are shown as blue squares ($\Delta h_{\text{avg}} = 0$) and green diamonds ($\Delta h_{\text{avg}} = 1$). The average and standard deviation of the Spearman correlation coefficient calculated for the 100 randomized happiness scores (null model) are shown as red circles with error bars (the error bars are smaller than the symbol). This data supports the hypothesis that happiness is less assortative as network distance increases.

Lastly, we explore whether these correlations are due to similarity of word usage. For this analysis, we compute the similarity of word bags for users connected in the reciprocal reply networks. We compare the distribution of observed similarity scores to similarity scores obtained by randomly reassigning word bags to users. Figure 2.A8 shows that both distributions are of a similar form, with the randomized version exhibiting a slightly lower mean similarity score ($\overline{D_{i,j}} = .167$) as compared to the mean of the observed similarity scores for users ($\overline{D_{i,j}} = .267$). If users were tweeting similar words with a similar frequency, we would expect a much larger mean similarity score for the observed data. Thus, we do not find evidence suggesting that the happiness correlations are due to similarity of word bags.

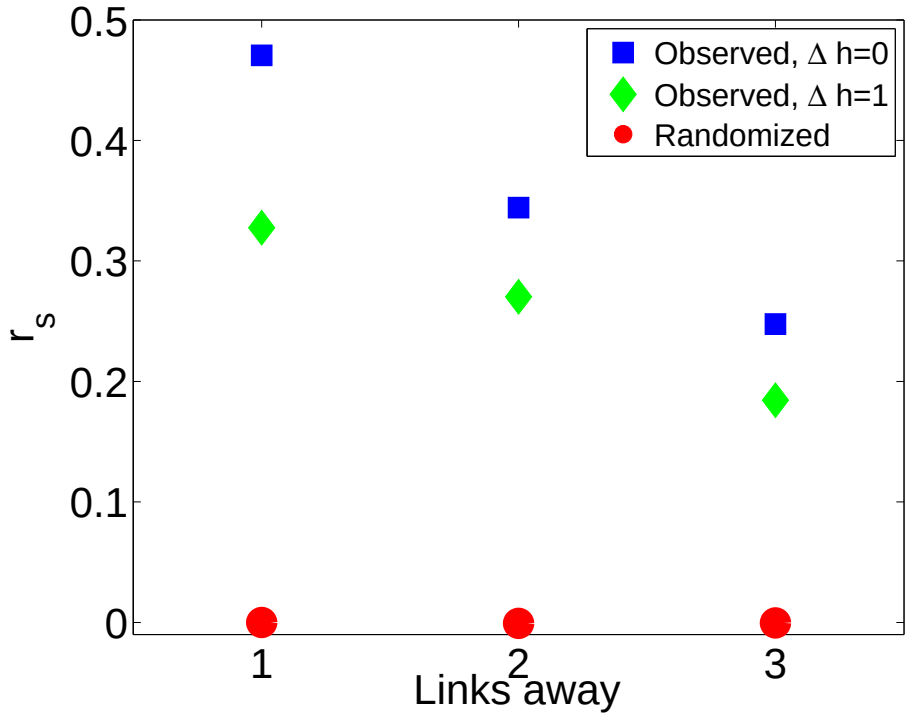


Figure 2.13: Application of null model to happiness scores. One hundred random permutations were applied to the happiness scores associated with each node in a sample week network (week beginning October 8, 2008 is shown), with $\Delta h = 0$ (blue square) and $\Delta h = 0$ (green diamonds). The threshold for all cases is set to $\alpha = 50$. The Spearman correlation coefficients, r_s for the observed data are shown as blue squares. The average and standard deviation of the Spearman correlation coefficient calculated for the 100 randomized data (null model) are shown as red circles with error bars (the error bars are smaller than the symbol). The plot shows Spearman correlation coefficients for the null model to be nearly 0 and provides supporting evidence for our observed trend, namely the network is assortative with respect to happiness and the strength of assortativity decreases as path length increases.

2.4 Discussion

In this paper, we describe how a social sub-network of Twitter can be derived from reciprocal-replies. Countering claims that Twitter is not social a network [15], we provide evidence of a very social Twitter. The large volume of replies (millions every week) and assortativity of user happiness indicates that Twitter is being used as a social service. Furthermore, conducted at the level of weeks, our analysis examines an in the moment social network, rather than the stale accumulation of social ties over a longer period of time. A network in which edges are created and never disintegrate results in dead links with no contemporary functional activity. This problem of unfriending has been noted [26] and can greatly impact conclusions drawn when observational data are used to infer contagion.

Our characterization of the reciprocal reply network reveals several trends over the 25 week period from September 2008 to February 2009. The number of nodes, N, in a given week network increased as time progressed, which is undoubtedly due to Twitter's enormous growth in popularity over the study period. Similarly, with an increasing number of nodes, we observe a smaller proportion of closed triangles (i.e., clustering shows a slight decrease). This may be due in part to sub-sampling effects or due to an increasing N, with which the number of closed triangles (i.e., friends of friends) cannot keep up. The proportion of nodes in the giant component remains fairly constant, as does degree assortativity as measured by Spearman's correlation coefficient. Had we used the Pearson correlation coefficient, degree assortativity would have been highly variable (Fig. A1) due to the extreme values of maximum degree ($k_{\max}$) during weeks 12-14 and 22. Using the Spearman rank correlation coefficient, which is less sensitive to extreme values, we find that the degree assortativity is fairly constant.

Our work is based on a sub-sample of tweets and is thus subject to the effects of missing data. The problem of missing data has been addressed by several researchers investigating the impact of missing nodes [43–47], missing links, or both [48]. More specifically, the work of Stumpf [43] shows that sub-sampled scale-free networks are not necessarily themselves scale-free. Further work which addresses the problem of missing messages and identifies the consequences of missing data on inferred network topology is needed to more fully address these questions.

We find support for the "happiness is assortative" hypothesis and evidence that these correlations can be detected up to three links away. Further, this finding does not appear to be based on users tweeting similar words (Fig. 2.A8). Our correlation coefficients for reciprocal-reply networks constructed at the level of weeks are smaller than those obtained by Bollen et al. [8] for a reciprocal-follower network constructed by aggregating over a six month period. This difference is likely a reflection of differences in methodologies, such as our more dynamic time scale (one-week periods vs. six month periods), our exclusion of central value happiness scores (i.e., stop words), and our use of the Spearman correlation coefficient.

While this paper does not attempt to separate homophily and contagion, future work could use reciprocal-reply networks to investigate these effects. While reciprocal-reply networks are subject to errors caused by missing data (see above discussion of this issue) they may provide a valuable framework for studying contagion effects, given that they are based on a conservative and dynamic metric of what constitutes an interaction on Twitter. A network structure in which links are known to be active and valid provides an arena in which the diffusion of information and emotion may be properly studied.

2.5 Acknowledgments

The authors acknowledge the Vermont Advanced Computing Core which is supported by NASA (NNX-08AO96G) at the University of Vermont for providing High Performance Computing resources that have contributed to the research results reported within this paper. CAB was supported by the UVM Complex Systems Center Fellowship Award, KDH was supported by VT NASA EPSCoR, and PSD was supported by NSF Career Award # 0846668. CMD and PSD were also supported by a grant from the MITRE Corporation.

2.6 References

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

[2] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A case study of unbiased sampling of OSNs. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9, 2010.

[3] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in Facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN '09, pages 37–42, New York, NY, USA, 2009.

[4] Z. Papacharissi. The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld. *New Media and Society*, 11(1-2):199–220, 2009.

[5] P. S. Dodds and C. M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11:441–456, 2010.

[6] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: An analysis of a microblogging community. In Haizheng Zhang, Myra Spiliopoulou, Bamshad Mobasher, C. Giles, Andrew McCallum, Olfa Nasraoui, Jaideep Srivastava, and John Yen, editors, *Advances in Web Mining and Web Usage Analysis*, volume 5439 of *Lecture Notes in Computer Science*, pages 118–138. Springer Berlin / Heidelberg, 2009.

[7] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everone's an influencer: Quantifying influence on Twitter. In *WSDM '11: Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, New York, NY, USA, 2011.

[8] J. Bollen, B. Goncalves, G. Ruan, and H. Mao. Happiness is assortative in online social networks. *Artificial Life*, 17(3), 2011.

[9] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[10] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in Twitter: The million follower fallacy. 2010.

[11] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS one*, 6(12):e26752, 2011.

[12] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. (E.) Zhao. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 369–378, New York, NY, USA, 2009.

[13] B. H. Huberman, D. H. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *CoRR*, abs/0812.1045, 2008.

[14] E. Kim, S. Gilbert, M. J. Edwards, and E. Graeff. Detecting sadness in 140 characters: Sentiment analysis and mourning Michael Jackson on Twitter. *Web Ecology*, 3, 2009.

[15] 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*, WWW '10, pages 591–600, New York, NY, USA, 2010.

[16] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[17] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010.

[18] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1397–1405. ACM, 2011.

[19] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012.

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

[21] J. H. Fowler and N. A Christakis. Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ*, 337, 2008.

[22] N. A. Christakis and J. H. Fowler. The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 358(21):2249–2258, 2008.

[23] J. N. Rosenquist, J. Murabito, J. H. Fowler, and N. A. Christakis. The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, 152(7):426–433, 2010.

[24] A. L. Hill, D. G. Rand, M. A. Nowak, and N. A. Christakis. Emotions as infectious diseases in a large social network: the SISa model. *Proceedings of the Royal Society B: Biological Sciences*, 277(1701):3827–3835, 2010.

[25] N. A. Christakis and J. H. Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*, 32:556–577, 2013.

[26] H. Noel, W. Galuba, and B. Nyhan. The unfriending problem: The consequences of homophily in friendship retention for causal estimates of social influence. *Social Networks*, 33:211–218, 2011.

[27] R. Lyons. The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy*, 2(1):1–26, 2011.

[28] C. R. Shalizi and A. C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, 40(2):211–239, 2011.

[29] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of World Wide Web Conference*, 2011.

[30] Twitter. Twitter API Blog. *http://blog.twitter.com/2011/09/one-hundred-million-voices*, 2011.

[31] B. Gonçalves, N. Perra, and A. Vespignani. Modeling users' activity on Twitter networks: Validation of Dunbar's Number. *PLoS one*, 6, 08 2011.

[32] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458, 2007.

[33] R. Grannis. Six degrees of "Who cares?". *American Journal of Sociology*, 115(4):991–1017, 2010.

[34] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science Magazine*, 333:1878–1881, 2011.

[35] G. Miller. Social scientists wade into the tweet stream. *Science Magazine*, 333:1814–1815, 2011.

[36] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy*, 98:404–409, 2001.

[37] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89:208701, 2002.

[38] I. M. Kloumann, C. M. Danforth, K. D. Harris, C. A. Bliss, and P. S. Dodds. Positivity of the English language. *PLoS one*, 7(1):e29484, 01 2012.

[39] M. Bastian, H. Sebastien, and J. Mathieu. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*, 2009.

[40] M. Jacomy, S. Heymann, T. Venturini, and M. Bastian. Forceatlas2, a graph layout algorithm for handy network visualization. `http://www.medialab.sciences-po.fr/publications/Jacomy_Heymann_Venturini-Force_Atlas2.pdf`, 2012.

[41] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[42] R. I. M. Dunbar. Neocortex size and group size in primates: A test of the hypothesis. *Journal of Human Evolution*, 28(3):287 – 296, 1995.

[43] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4221–4224, 2005.

[44] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 55–64, New York, NY, USA, 2011. ACM.

[45] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 631–636, New York, NY, USA, 2006. ACM.

[46] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(1):016102, 2006.

[47] T.L. Frantz, M. Cataldo, and K.M. Carley. Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328, 2009.

[48] G. Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3):247–268, 2006.

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

2.7 Appendix

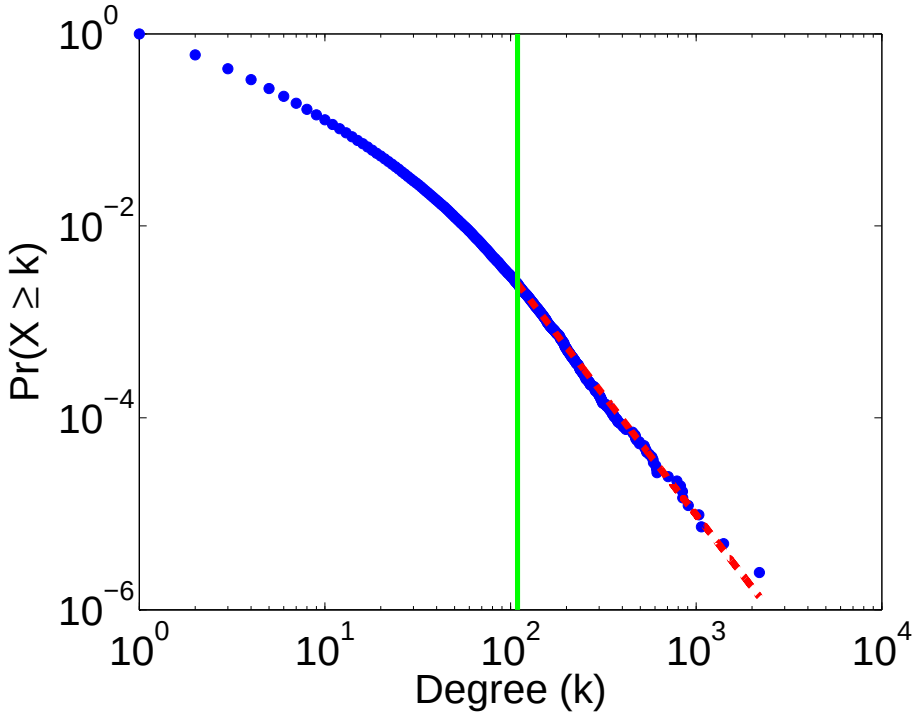


Figure 2.A1: Degree distribution for a sample week. Log-log plot of the complementary cumulative distribution function (CCDF) of the degree distribution for a sample month (January 2009) network is shown (blue), along with the best fitting power law model ($\alpha = 3.50$ and $k_{\min} = 109$) using the procedure of Clauset, Shalizi, and Newman [41]. We test whether the empirical distribution is distinguishable from a power law using the Kolmogorov-Smirnov test and find no evidence against the null hypothesis ($D = 1.82 \times 10^{-2}, p = 0.35, n = 495881$) data. This distribution may be fit by a power law.

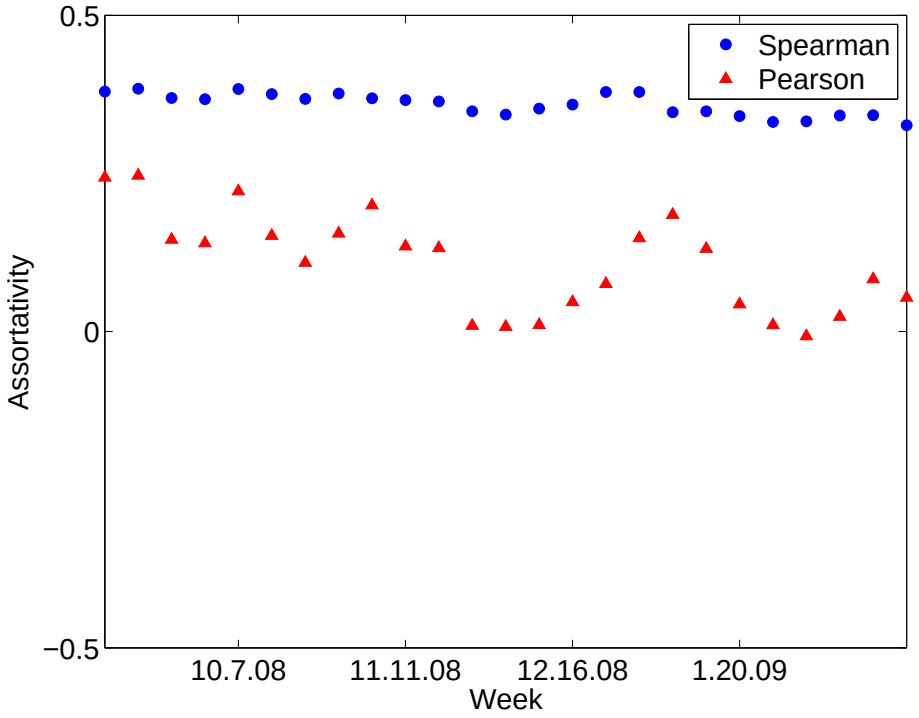
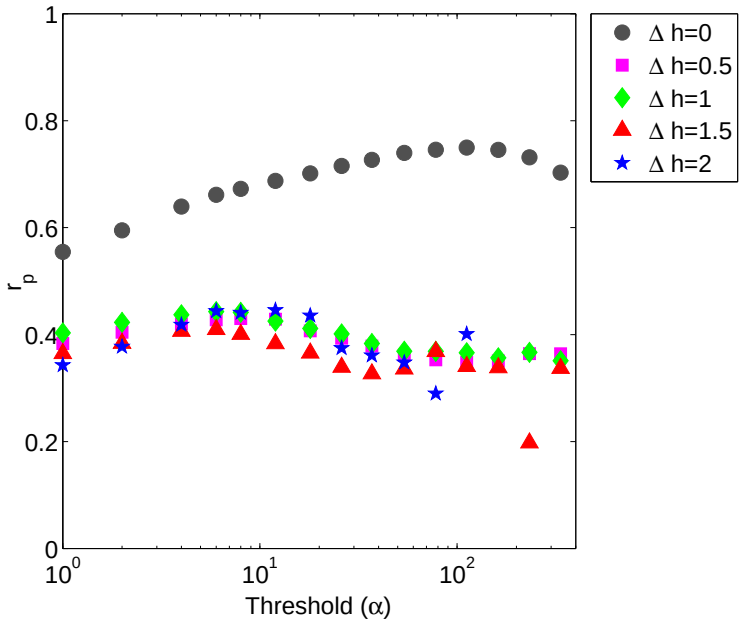
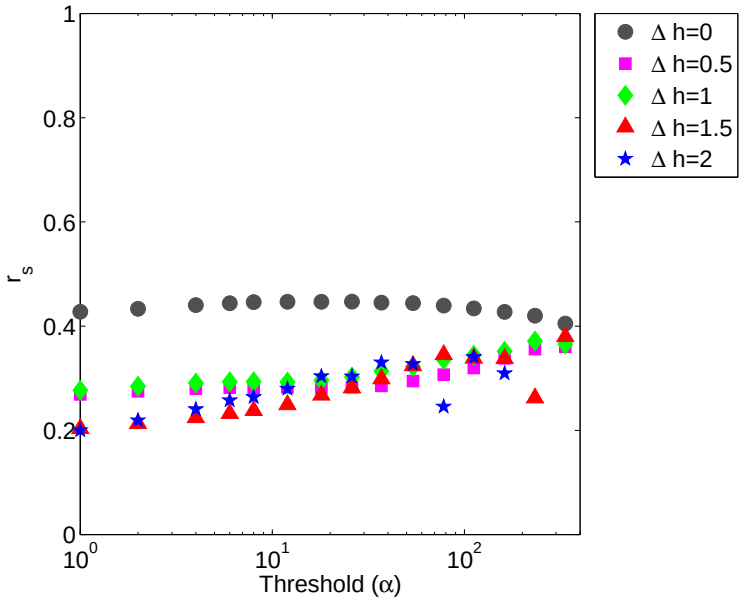


Figure 2.A2: Comparison of Spearman and Pearson correlation coefficients for assortativity. Spearman and Pearson correlation coefficients are used to measure degree assortativity. The Pearson correlation coefficient is more sensitive to extreme values. As a result, the Pearson correlation coefficient obscures the trend that the network is assortative with respect to the rank of node degrees. Given the nature of the degree distribution and the questions that we are asking, we use the Spearman correlation coefficient for our study.



(a) Assortativity of happiness, Pearson's r



(b) Assortativity of happiness, Spearman's r

Figure 2.A3: Happiness assortativity vs. word count threshold. Measured happiness assortativity as threshold for labMT word usage increases for a single week network. The Spearman correlation coefficient (right) exhibits less variability as compared to the Pearson correlation coefficient (left). Notice that when $\Delta h = 0$, there is less variation due to the numerous words centered around the mean happiness score, regardless of the threshold, α.

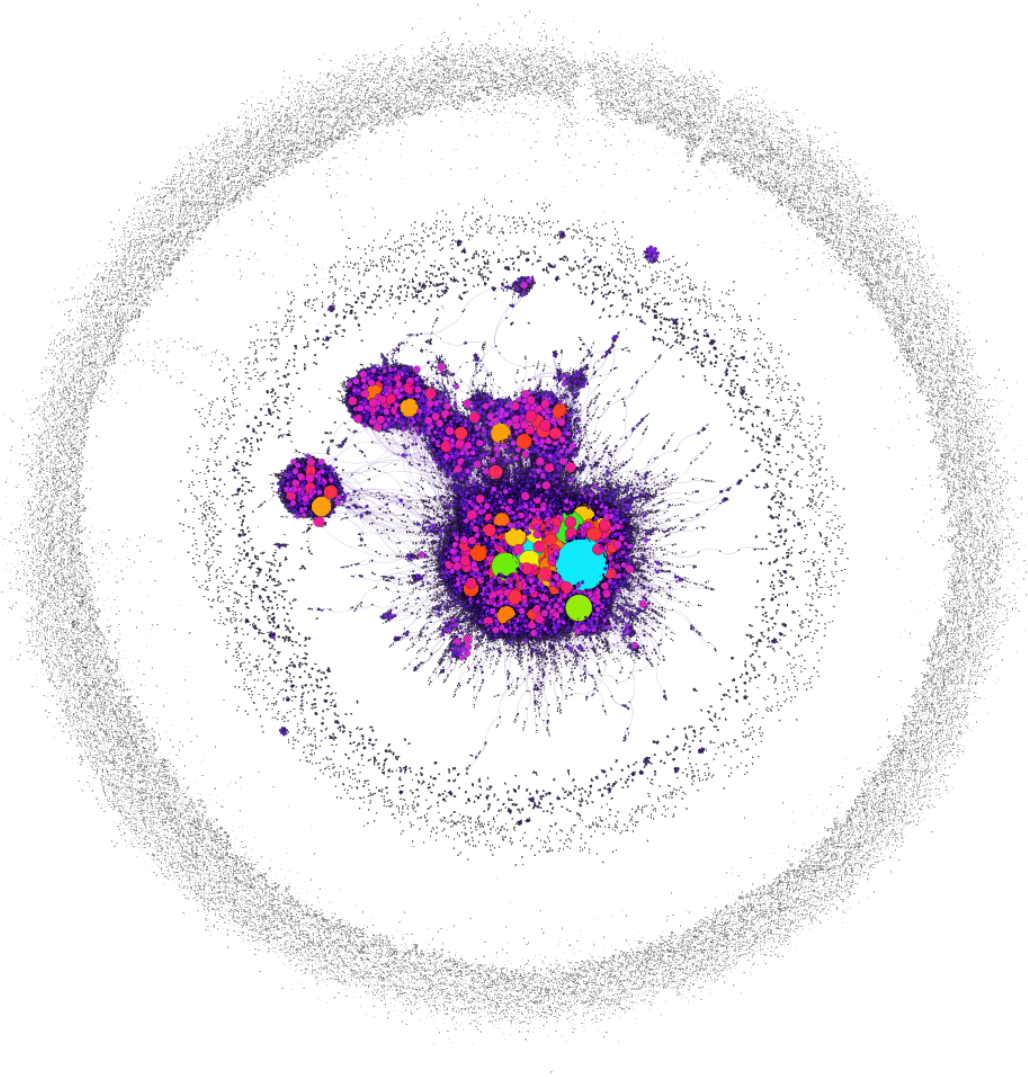


Figure 2.A4: Visualization of a reciprocal reply network with emphasis on degree. A visualization of the reciprocal reply network for the week beginning September 9, 2008 (Week 1) is depicted. The size of a node is proportional to the degree, and colors further emphasize the degree detected by Gephis implementation of the algorithm suggested by Blondel et al. [40].

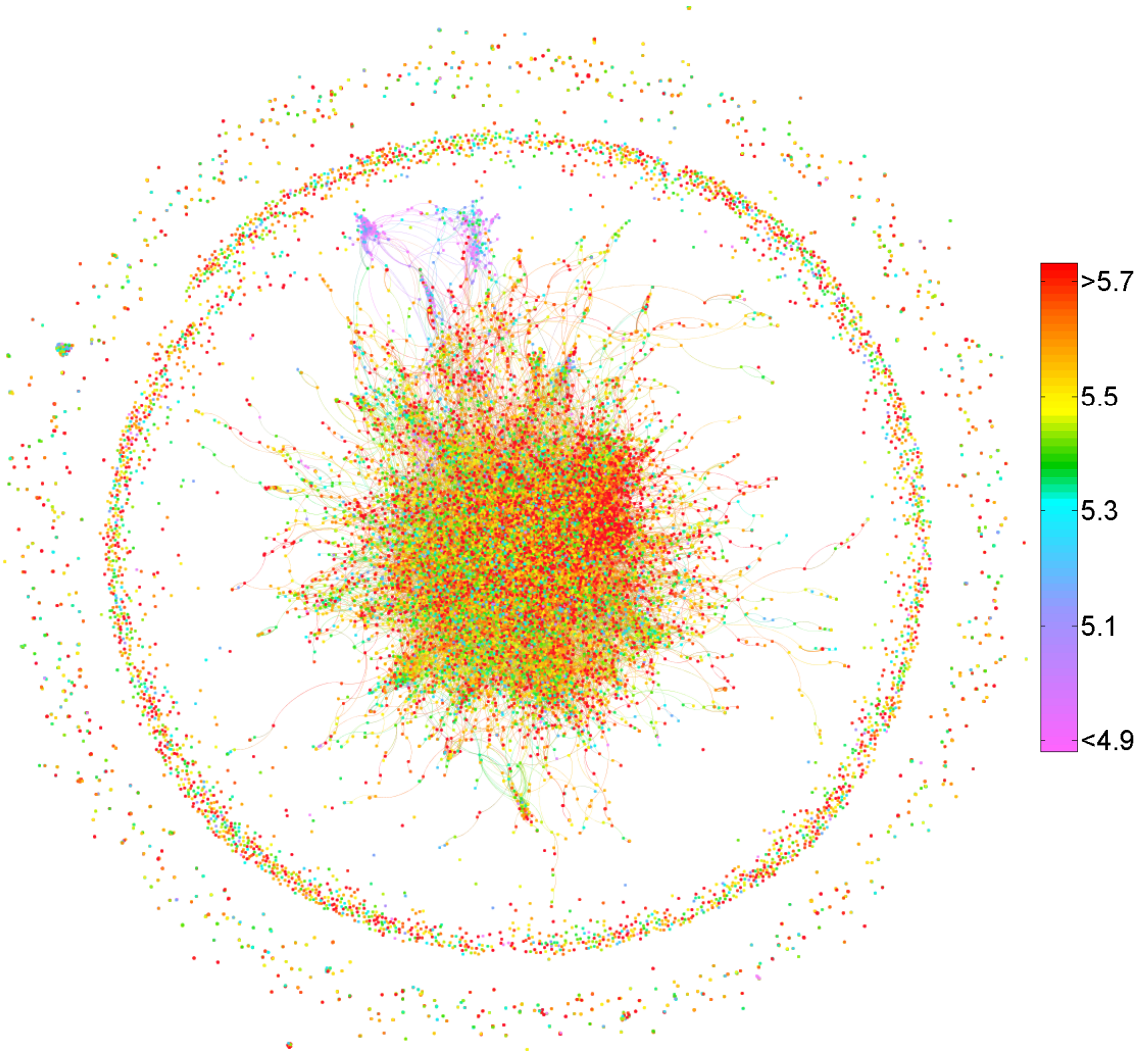


Figure 2.A5: Visualization of the reciprocal reply network for the week beginning September 9, 2008 (Week 1) where colors represent happiness scores for nodes with greater than $\alpha = 50$ labMT words (57% of all nodes in the week). The visualization was produced using Gephi [39]. The algorithm employed by the software clusters nodes according to their connectivity. Collections of nodes with similar colors provide a visualization of the happiness is assortativity finding.

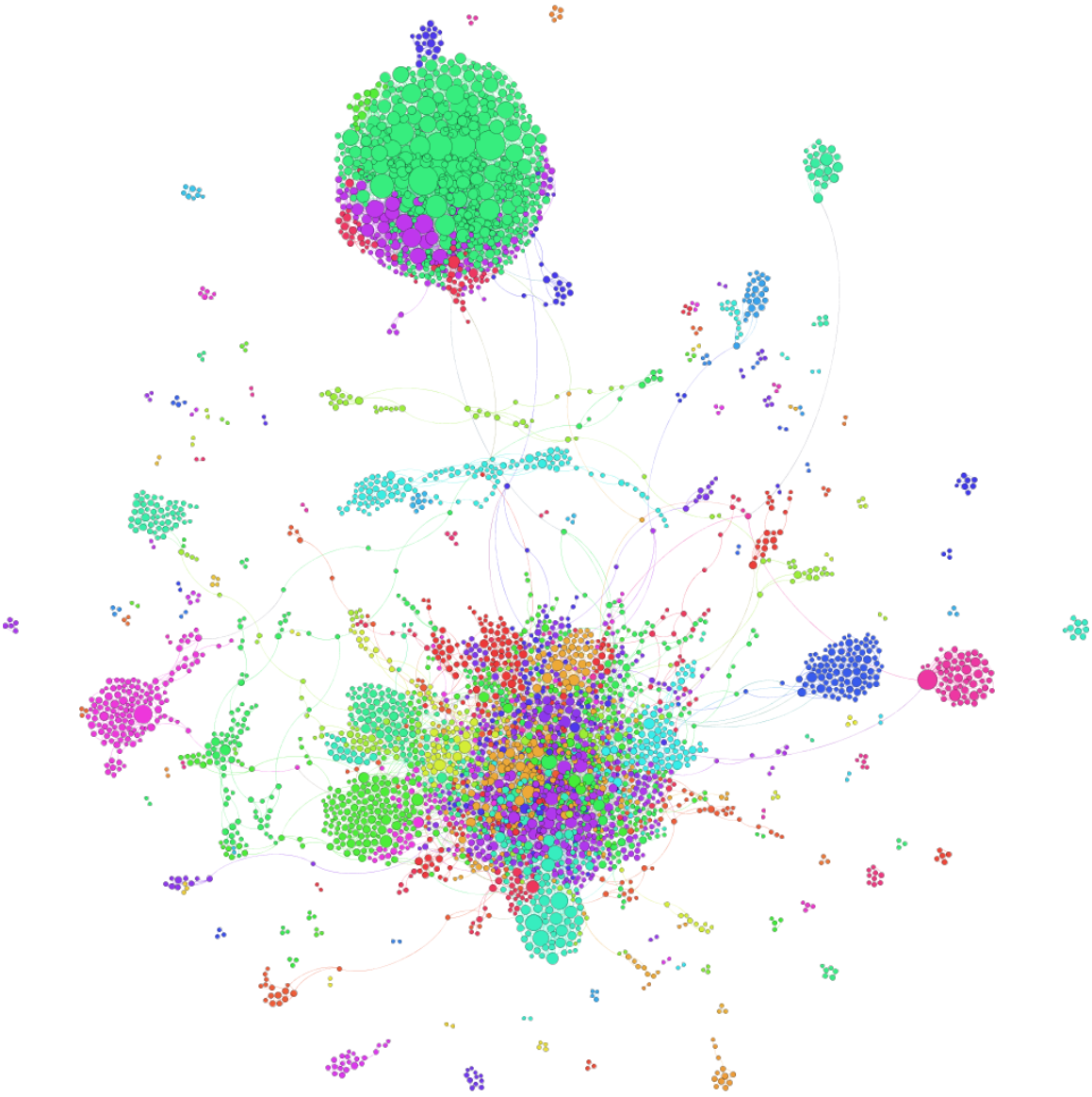


Figure 2.A6: A visualization of the reciprocal reply network for the day of October 28, 2008. The size of the nodes is proportional to the degree and colors indicate communities detected by Gephi's implementation of the community detection algorithm suggested by [49].

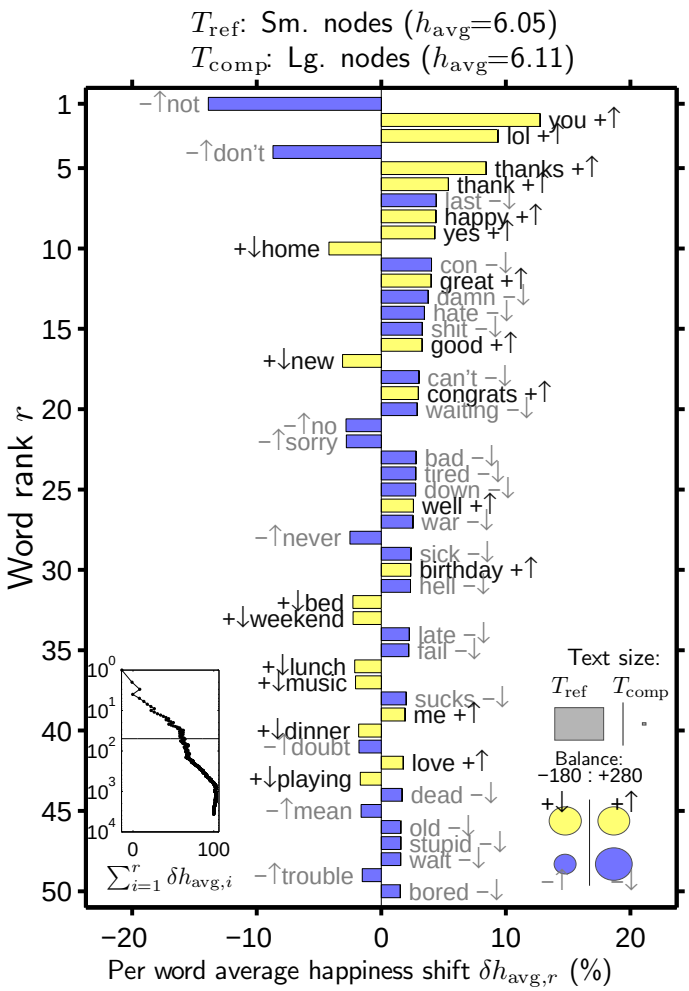


Figure 2.A7: Wordshift for large and small degree nodes. The collection of words used by nodes $k \geq 100$ ($T_{\rm comp}$) is compared to words written by users $k < 100$ ($T_{\rm ref}$). The horizontal bars on the right side of the plot represent words which raise the happiness score of $T_{\rm comp}$. The symbols of $+/-$ and $\uparrow$ / $\downarrow$ combine to convey whether a positive/negative word appears more/less frequently in the $T_{\rm comp}$ as compared to the $T_{\rm ref}$. Notice that an increase in the usage of positive words (e.g., "you"), as well as a decrease in the use of a negative word (e.g., "last") will contribute to $T_{\rm comp}$ having a higher happiness score. In the lower right, the relative text sizes are depicted as rectangles proportional to the number of words. The circle plots depicted the relative amount of positive vs. negative words contained in $T_{\rm ref}$ and $T_{\rm comp}$. While both collections are similar in terms of positive word usage, the collection of words used by larger nodes contains fewer negative words and thus, this contributes to the slightly higher happiness score for this collection of words. The lower left inset shows the cumulative sum of individual word contributions as a function of $\log_{10} r$, where r is the rank of the 3,686 labMT words. See [11] for the full details of the wordshift graph.

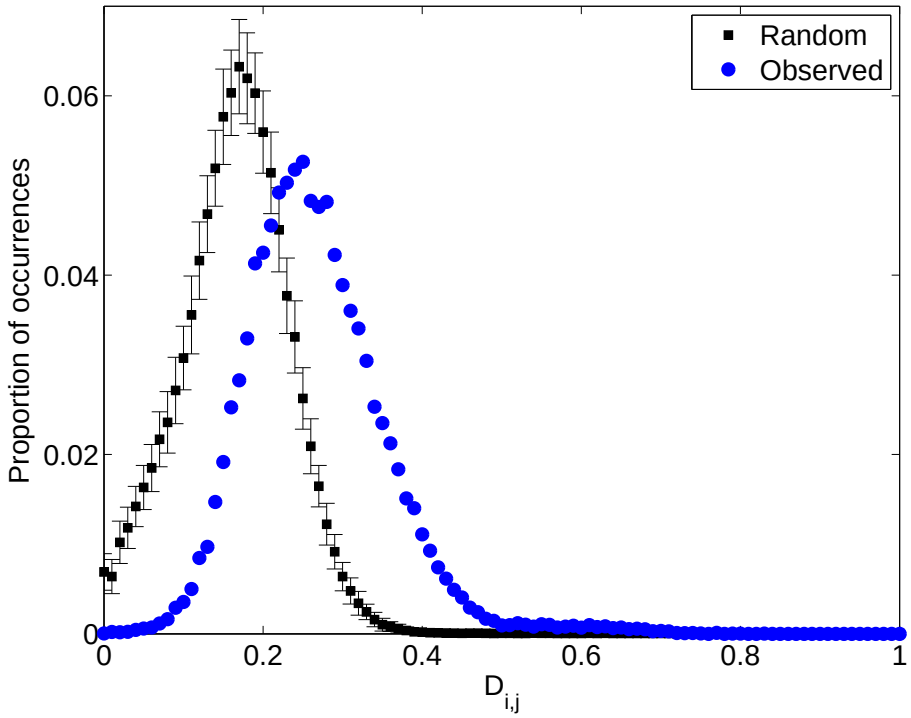


Figure 2.A8: Similarity scores of word bags and null model. The similarity of word bags for pairs of users connected in a week reciprocal reply network is computed as follows: For users i and j, we compute $D_{i,j} = 1 - \frac{1}{2}\sum_{n=1}^{3686} |f_{i,n} - f_{j,n}|$, where $f_{i,n}$ represents the normalized frequency of word usage of the nth labMT word by user i. The value of $D_{i,j}$ ranges from 0 (dissimilar word bags) to 1 (similar word bags). The proportion of occurrences of user-user pairs in the reciprocal reply network for a sample week (Sept. 16, 2008) having word similarity indices between 0 and 1 are shown (blue dots), with $\alpha = 50$ and $\Delta h = 1$. The majority of user-user similarity indices are less than 0.4, indicating that users and their nearest neighbors use dissimilar collections of words in their tweets. We then perform 100 random permutations of word vector assignments to users, while holding the network topology intact (black squares). The resulting distributions show that while users are using more similar words than would be expected by chance, this shift is small. The mean score for randomized user-user paired word collections is $\overline{D_{i,j}} = .167$. This value is not zero, since users are using a common language (English). The mean score for our observed network data is $\overline{D_{i,j}} = .267$, which is slightly higher than the randomized value due to conversations occurring between these users.

Rank	Word	Frequency ($\times 10^5$)	Happiness	Rank	Word	Frequency ($\times 10^5$)	Happiness	Rank	Word	Frequency ($\times 10^5$)	Happiness
1	you	103.55	6.24	41	happy	8.40	8.30	81	google	5.05	7.20
2	my	94.91	6.16	42	tomorrow	7.88	6.18	82	everyone	5.03	6.12
3	me	56.35	6.58	43	nice	7.80	7.38	83	most	4.95	6.22
4	not	39.98	3.86	44	best	7.61	7.18	84	wait	4.88	3.74
5	up	36.04	6.14	45	she	7.57	6.18	85	start	4.87	6.10
6	no	34.40	3.48	46	yes	7.42	6.74	86	please	4.79	6.36
7	new	34.03	6.82	47	fun	7.37	7.96	87	con	4.78	3.70
8	like	31.75	7.22	48	hope	7.34	7.38	88	try	4.77	6.02
9	all	30.71	6.22	49	bad	6.98	2.64	89	thought	4.69	6.38
10	good	30.20	7.20	50	never	6.92	3.34	90	school	4.66	6.26
11	will	23.58	6.02	51	sure	6.82	6.32	91	thank	4.64	7.40
12	we	22.59	6.38	52	done	6.81	6.54	92	weekend	4.56	8.00
13	day	21.80	6.24	53	show	6.73	6.24	93	hey	4.48	6.06
14	know	19.45	6.10	54	awesome	6.72	7.60	94	wish	4.44	6.92
15	more	19.32	6.24	55	check	6.51	6.10	95	hate	4.42	2.34
16	don't	18.29	3.70	56	bed	6.42	7.18	96	haha	4.41	7.64
17	today	18.24	6.22	57	sleep	6.33	7.16	97	friends	4.41	7.92
18	love	17.66	8.42	58	cool	6.32	7.20	98	making	4.40	6.24
19	think	17.45	6.20	59	live	6.28	6.84	99	dinner	4.27	7.40
20	see	15.28	6.06	60	big	6.28	6.22	100	coffee	4.27	7.18
21	great	14.60	7.88	61	free	6.18	7.96	101	music	4.24	8.02
22	lol	13.35	6.84	62	life	6.17	7.32	102	found	4.23	6.54
23	thanks	13.09	7.40	63	old	6.07	3.98	103	doesn't	4.23	3.62
24	home	13.05	7.14	64	didn't	6.04	4.00	104	online	4.23	6.72
25	people	12.71	6.16	65	find	6.00	6.00	105	party	4.20	6.34
26	night	12.70	6.22	66	die	6.00	1.74	106	soon	4.20	6.34
27	blog	12.26	6.02	67	video	5.99	6.48	107	thinking	4.15	6.28
28	last	11.89	3.74	68	house	5.99	6.34	108	snow	4.14	6.32
29	well	11.70	6.68	69	christmas	5.89	7.96	109	give	4.13	6.54
30	make	11.27	6.00	70	playing	5.77	7.14	110	movie	4.12	6.84
31	right	11.04	6.54	71	world	5.76	6.52	111	ha	4.09	6.00
32	can't	10.93	3.42	72	game	5.54	6.92	112	sorry	4.08	3.66
33	morning	10.38	6.56	73	wow	5.54	7.46	113	real	4.06	6.78
34	very	10.10	6.12	74	ready	5.53	6.58	114	kids	3.98	7.38
35	first	9.69	6.82	75	iphone	5.53	6.54	115	phone	3.91	6.44
36	our	9.26	6.08	76	listening	5.41	6.28	116	tv	3.91	6.70
37	better	8.89	7.00	77	pretty	5.40	7.32	117	stop	3.89	3.90
38	us	8.82	6.26	78	always	5.39	6.48	118	play	3.88	7.26
39	tonight	8.79	6.14	79	help	5.27	6.08	119	waiting	3.88	3.68
40	down	8.73	3.66	80	read	5.07	6.52	120	lunch	3.81	7.42

Table 2.A1: Most frequently occurring words (stop words removed). The top 120 most frequently occurring words from the labMT list in our Sept 2008 through Feb 2009 data set, where stop words ($4 < h_{\text{avg}} < 6$) have been removed.

Rank	Word	Frequency ($\times 10^5$)	Happiness	Rank	Word	Frequency ($\times 10^5$)	Happiness	Rank	Word	Frequency ($\times 10^5$)	Happiness
1	the	295.60	4.98	41	what	29.46	4.80	81	off	14.89	4.02
2	to	249.91	4.98	42	about	28.97	5.16	82	great	14.60	7.88
3	i	221.28	5.92	43	it's	27.14	4.88	83	need	14.45	4.84
4	a	218.13	5.24	44	if	25.21	4.66	84	he	14.34	5.42
5	and	135.23	5.22	45	by	24.66	4.98	85	still	13.74	5.14
6	is	127.94	5.18	46	as	24.50	5.22	86	been	13.43	5.04
7	in	122.94	5.50	47	time	24.19	5.74	87	lol	13.35	6.84
8	of	121.79	4.94	48	one	23.73	5.40	88	would	13.15	5.38
9	for	114.41	5.22	49	will	23.58	6.02	89	thanks	13.09	7.40
10	you	103.55	6.24	50	can	23.57	5.62	90	home	13.05	7.14
11	on	96.97	5.56	51	an	22.73	4.84	91	want	12.81	5.70
12	my	94.91	6.16	52	we	22.59	6.38	92	people	12.71	6.16
13	it	91.09	5.02	53	some	22.32	5.02	93	night	12.70	6.22
14	that	69.81	4.94	54	que	22.26	4.64	94	here	12.28	5.48
15	at	58.51	4.90	55	day	21.80	6.24	95	o	12.26	4.96
16	with	56.42	5.72	56	how	21.64	4.68	96	blog	12.26	6.02
17	me	56.35	6.58	57	going	20.64	5.42	97	why	12.10	4.98
18	just	50.25	5.76	58	am	20.60	5.38	98	much	11.92	5.74
19	have	49.86	5.82	59	go	20.03	5.54	99	last	11.89	3.74
20	be	46.10	5.68	60	has	19.68	5.18	100	did	11.84	5.58
21	this	45.75	5.06	61	or	19.55	4.98	101	el	11.76	4.80
22	de	44.38	4.82	62	know	19.45	6.10	102	well	11.70	6.68
23	so	40.93	5.08	63	more	19.32	6.24	103	oh	11.69	4.84
24	not	39.98	3.86	64	la	18.77	5.00	104	who	11.64	5.06
25	i'm	39.89	5.74	65	don't	18.29	3.70	105	should	11.48	5.24
26	are	39.03	5.16	66	today	18.24	6.22	106	over	11.34	4.82
27	but	37.78	4.24	67	too	18.15	5.22	107	make	11.27	6.00
28	was	37.74	4.60	68	they	18.09	5.62	108	then	11.15	5.34
29	up	36.04	6.14	69	work	17.95	5.24	109	right	11.04	6.54
30	out	35.20	4.62	70	got	17.91	5.60	110	can't	10.93	3.42
31	now	35.12	5.90	71	love	17.66	8.42	111	way	10.84	5.24
32	no	34.40	3.48	72	think	17.45	6.20	112	only	10.72	4.92
33	new	34.03	6.82	73	back	17.37	5.18	113	getting	10.63	5.68
34	do	33.96	5.76	74	twitter	17.18	5.46	114	his	10.56	5.56
35	from	33.78	5.18	75	when	16.84	4.96	115	morning	10.38	6.56
36	like	31.75	7.22	76	there	16.39	5.10	116	very	10.10	6.12
37	your	31.43	5.60	77	had	15.30	4.74	117	after	9.82	5.08
38	all	30.71	6.22	78	see	15.28	6.06	118	watching	9.76	5.84
39	good	30.20	7.20	79	en	14.97	4.84	119	her	9.73	5.84
40	get	30.04	5.92	80	really	14.93	5.84	120	them	9.71	4.92

Table 2.A2: Top 120 most frequently occurring words from the labMT word list in our Sept 2008 through Feb 2009 data set including stop words.

Week	Start date	N	$< k >$	$k_{\max}$	C_G	Assort	# Comp.	S
1	09.09.08	95647	2.99	261	0.10	0.24	10364	0.71
2	09.16.08	99236	2.95	313	0.10	0.24	11062	0.71
3	09.23.08	99694	2.90	369	0.09	0.13	11457	0.70
4	09.30.08	100228	2.87	338	0.09	0.13	11752	0.69
5	10.07.08	78296	2.60	241	0.09	0.21	11140	0.63
6	10.14.08	122644	3.20	394	0.09	0.14	12221	0.74
7	10.21.08	130027	3.30	559	0.08	0.09	12420	0.75
8	10.28.08	144036	3.56	492	0.08	0.14	12319	0.78
9	11.04.08	145346	3.54	330	0.08	0.19	12597	0.78
10	11.11.08	136534	3.35	441	0.08	0.12	12972	0.76
11	11.18.08	153486	3.46	444	0.08	0.13	13594	0.77
12	11.25.08	155753	3.46	1244	0.06	0.00	14122	0.77
13	12.02.08	165156	3.44	1245	0.06	0.01	14496	0.78
14	12.09.08	162445	3.33	1456	0.05	0.01	15342	0.76
15	12.16.08	148154	3.12	730	0.06	0.04	15645	0.73
16	12.23.08	140871	3.22	575	0.07	0.07	15216	0.72
17	12.30.08	143015	3.30	519	0.07	0.15	15272	0.73
18	01.06.09	170597	3.19	253	0.07	0.18	17234	0.74
19	01.13.09	188429	3.29	477	0.07	0.13	18403	0.75
20	01.20.09	196038	3.16	680	0.06	0.04	19927	0.74
21	01.27.09	203852	3.04	973	0.05	0.01	21537	0.73
22	02.03.09	212513	2.92	1718	0.04	-0.01	24387	0.71
23	02.10.09	213936	2.83	828	0.06	0.02	25854	0.70
24	02.17.09	215172	2.65	437	0.06	0.07	28742	0.67
25	02.24.09	170180	2.27	320	0.06	0.04	28388	0.58

Table 2.A3: Network statistics for reciprocal-reply networks by week. As Twitter popularity grows, so does the number of users (N) in the observed reciprocal-reply network. The average degree ($< k >$), degree assortativity, the number of nodes in the giant component (# Comp.), and the proportion of nodes in the giant component (S) remain fairly constant, whereas the maximum degree ($k_{\max}$) shows a great deal of variability from month to month. Clustering (C_G) shows a slight decrease over the course of this period.

Week	Start date	# Obsvd. Msgs. $\times 10^6$	# Total Msgs. $\times 10^6$	% Obsvd.	# Replies $\times 10^6$	% Replies
1	09.09.08	3.14	7.26	43.2	0.88	28.1
2	09.16.08	3.36	8.31	40.4	0.90	26.9
3	09.23.08	3.43	8.89	38.6	0.90	26.2
4	09.30.08	3.33	9.06	36.8	0.89	26.6
5	10.07.08	2.33	9.38	24.8	0.64	27.5
6	10.14.08	4.39	9.87	44.4	1.24	28.3
7	10.21.08	4.70	10.01	47.0	1.35	28.8
8	10.28.08	5.74	10.34	55.5	1.64	28.5
9	11.04.08	5.58	11.14	50.1	1.63	29.3
10	11.11.08	4.70	9.88	47.6	1.42	30.2
11	11.18.08	5.48	11.34	48.3	1.67	30.5
12	11.25.08	5.71	11.47	49.8	1.73	30.2
13	12.02.08	5.54	12.85	43.1	1.80	32.4
14	12.09.08	5.41	13.54	39.9	1.72	31.7
15	12.16.08	4.57	12.72	35.9	1.45	31.8
16	12.23.08	4.80	11.62	41.3	1.46	30.5
17	12.30.08	4.61	13.48	34.2	1.50	32.5
18	01.06.09	5.16	16.11	32.0	1.72	33.3
19	01.13.09	5.73	17.33	33.1	1.97	34.4
20	01.20.09	5.82	18.87	30.9	1.98	34.1
21	01.27.09	5.75	20.79	27.6	1.98	34.5
22	02.03.09	5.78	22.42	25.8	2.01	34.8
23	02.10.09	5.66	23.39	24.2	1.99	35.1
24	02.17.09	5.43	25.71	21.1	1.91	35.1
25	02.24.09	3.80	20.75	18.3	1.34	35.1

Table 2.A4: Number of observed messages in our database (September 2008 through February 2009). The number of "observed" messages in our database comprise a fraction of the total number of Twitter messages made during period of this study (September 2008 through February 2009). While our feed from the Twitter API remains fairly constant, the total # of tweets grows, thus reducing the % of all tweets observed in our database. We calculate the total # of messages as the difference between the last message id and the first message id that we observe for a given month. This provides a reasonable estimation of the number of tweets made per month as message ids were assigned (by Twitter) sequentially during the time period of this study. We also report the number observed messages that are replies to specific messages and the percentage of our observed messages which constitute replies.

Chapter 3

An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks

Many real world, complex phenomena have underlying structures of evolving networks where nodes and links are added and removed over time. A central scientific challenge is the description and explanation of network dynamics, with a key test being the prediction of short and long term changes. For the problem of short-term link prediction, existing methods attempt to determine neighborhood metrics that correlate with the appearance of a link in the next observation period. Recent work has suggested that the incorporation of topological features and node attributes can improve link prediction. We provide an approach to predicting future links by applying the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) to optimize weights which are used in a linear combination of sixteen neighborhood and node similarity indices. We examine a large dynamic social network with over 10^6 nodes (Twitter reciprocal reply networks), both as a test of our general method and as a problem of scientific interest in itself. Our method exhibits fast convergence and high levels of precision for the top twenty predicted links. Based on our findings, we suggest possible factors which may be driving the evolution of Twitter reciprocal reply networks.

3.1 Introduction

Time varying social networks can be used to model groups whose dynamics change over time. Individuals, represented by nodes, may enter or exit the network, while interactions, represented by links, may strengthen or weaken. Most network growth models capture global properties, but do not capture specific localized dynamics such as who will be connected to whom in the future. And yet, it is precisely this type of information that would be most valuable in applications such as national security, online social networking sites (people you may know), and organizational studies (predicting potential collaborators).

In this paper, we focus primarily on the link prediction problem: given a snapshot of a network $G_t = (V, E_t)$, with nodes V (nodes present across all time steps) and links E_t, at time t, we seek to predict the most likely links to newly occur in the next timestep, $t+1$ [1].

Link prediction strategies may be broadly categorized into three groups: similarity based strategies, maximum likelihood algorithms, and probabilistic models. As noted by Lu et al. [2], the latter two approaches can be prohibitively time consuming for a large network over $10,000$ nodes. Given our interest in large, sparse networks with $N \gtrsim 10^6$, we focus primarily on local information and use similarity indices to characterize the likelihood of future interactions. We consider the two major classes of similarity indices: topological-based and node attribute (Table 3.1).

There does not appear to be one best similarity index that is superior in all settings. Depending on the network under analysis, various measures have shown to be particularly promising [1, 3–8]. These findings suggest that the predictors which work "best" for a given network may be related to the inherent structure within the individual network rather than a

universal best set of predictors. Further, it is also plausible that the best link predictor may change as the network responds to endogenous and exogenous factors driving its evolution.

Topological similarity indices encode information about the relative overlap between nodes' neighborhoods. We expect that the more "similar" two nodes' topological neighborhoods are (e.g., the more overlap in their shared friends), the more likely they may be to exhibit a future link. The common neighbors index, a building block of many other topological similarity indices, has been shown to correlate with the occurrence of future links [9]. Several variants of this index have been proposed and have been shown to be useful for link prediction in a variety of settings [3, 10–18]. See [2] for a review. In their seminal paper on link prediction, Liben-Nowell and Kleinberg [1] examined author collaboration networks derived from arXiv submissions in four subfields of Physics. They found that neighborhood similarity measures, such as the Jaccard [15], Adamic-Adar [19], and the Katz coefficients [20] provided a large factor improvement over randomly predicted links.

As a complement for topological similarity indices, node-specific similarity indices examine node attributes, such as language, topical similarity, and behavior, in the case of social networks. Several studies have suggested that incorporating these measures can enhance link prediction [2, 4, 22–26]. In training algorithms for link prediction, researchers have used supervised learning including support vector machine [27], decision trees [4], bagged random forests [17], supervised random walks [6], multi-layer perceptrons, and others. Notably, Al Hasan et al. [27] use both topological and node-specific features to compare several supervised learning algorithms. They found that support vector machine (SVM) performed the best for the prediction of future links. While SVM is often considered a state of the art supervised learning model, one of its major drawbacks relates to kernel

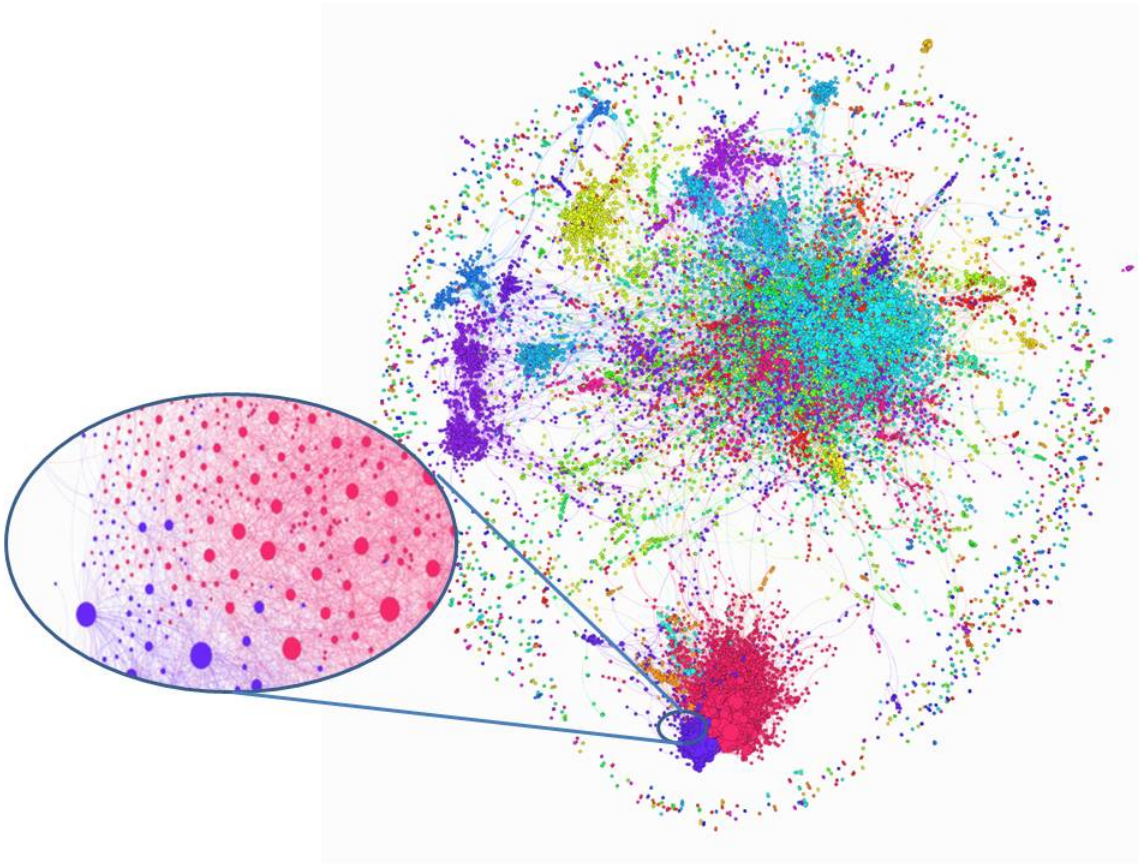


Figure 3.1: Visualization of persistent individuals and their interactions in a one week Twitter RRN. A visualization of a one week Twitter reciprocal reply network exhibiting interactions between a core of 25,936 users who were active in each of networks in the period from September 9, 2008 to October 20, 2008 reveals the large degree observed in one community (inset). The colors indicate modularity, a proxy for community structure, as detected by Gephi's implementation of Blondel's "Fast unfolding of communities in large networks" [21].

selection [28]. Furthermore, Litchenwalter et al. [17], who use Weka's implementation of bagged random forests to produce ensembles of models and reduce variance, note the need to undersample due to the computational complexity of their method on large datasets. Of particular interest, Wang et al. [4] study a network of individuals constructed from mobile phone call data. They compare similarity indices used in isolation to a link predictor com-

bining several indices (binary decision tree determined from supervised learning). These researchers found that the combination of node-specific and topological similarity indices outperform topological indices in isolation. While their results are promising, they acknowledge that the cost comes from looking at only a subset (e.g., 300 potential links which have Adamic-Adar scores > 0.5 and Spatial Co-location rate > 0.7) from the large potential set of user-user pairs two-links away (e.g., 266,750).

Motivated by the above, we aim here to provide a link predictor encompassing both topological and user-specific information, which exhibits fast convergence and which does not require parametric thresholds nor undersampling due to computational complexity.

In this paper, we fix a linear model for combining neighborhood similarity measures and node specific data and use an evolutionary algorithm to find the coefficients which optimize the proportion of correctly predicted links. Rather than pre-supposing that all similarity indices are of equal importance, we allow the weights of this linear combination to adjust using Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [29]). Clearly, the optimal model combining similarity indices may not be linear and our assumption of this model structure is a limitation of our work. With that said, our work has several advantages over other methods for link prediction and our work reveals that a simple, linear model produces comparable results (if not better), with the added advantage of suggesting possible mechanisms driving the network's evolution over time.

In many supervised learning approaches, link prediction efforts fit both a model structure and parameters. To surmount the challenge of large feature sets and large networks, researchers limit which features to include or perform undersampling due to computational complexity of these algorithms. Our approach of using CMA-ES for link prediction liberates researchers to include several indices in the link predictor, irrespective of their assumed

performance. This is a strength of our method in that no assumption of network class nor prior knowledge about the system under analysis is required.

Although we focus on the link prediction problem for a large, dynamic social network, our methods are independent of network type and may be applied to various biological, infrastructure, social and virtual networks. We demonstrate sixteen commonly used similarity indices here, but we emphasize that any other similarity indices may be interchanged for or added to the ones included in this study. The choice of which similarity measures to include will largely depend on available data (e.g., metadata for nodes and appropriate topological indices one has available in the context of the network one is studying) and the size of the network under consideration.

Another limitation of several supervised learning approaches for link prediction is that the interpretation of the model may yield little information about the the network's evolutionary processes. Our methods provide transparency and the detection of indices which function as good predictors for future links which can help to elucidate possible mechanisms which may be driving the evolution of the network over time.

In recent years, there has been a surge of interest in viewing Twitter activity through the lens of social network analysis. In many studies, nodes represent individuals and links represent following behavior [30–32], reciprocated following [33], replies [25] or reciprocated replies [34].

Our application will be link prediction in Twitter reciprocal reply networks (RRNs), a construction first proposed by Bliss et al. [34]. We examine the evolution of these networks constructed at the time scale of weeks, where nodes represent users and links represent evidence of reciprocated replies during the time period of analysis. While many other

studies have examined following and reciprocated following, we use reciprocated replies as evidence of social interaction and active engagement of individuals.[1]

Due to the large size of networks that we seek to study and the hypothesis that friends of friends are more likely to become friends than individuals who have no friends in common [35, 36], we restrict out attention to the prediction of new links at time $t + 1$ which occur between individuals who were separated by a path length of 2 at time t (i.e., triadic closure). Empirical evidence suggests that a preponderance of new links form between such 2-link neighbors in email reply networks [37], Twitter follower networks [38], and Twitter RRNs.[2]

Previous link prediction efforts related to Twitter have largely focused on predicting follower relationships. Rowe, Stankovic and Alani [23] use supervised learning to combine topological and node specific features (e.g., topics of tweets, tweet counts, re-tweets, etc.) to predict following behavior. Romero and Kleinberg also examined link prediction in follower networks and suggest that directed closure plays an important role in the formation of new links [38]. Hutto, Yardi, and Gilbert [24] examine 507 individuals and their followers to find that user-specific characteristics, such as message content and behavior should be given equal weight as topological characteristics for link prediction. Yin, Hong, and Davison examine 979 individuals and their neighbors (in Twitter follower networks) to predict following behavior over a six week time-scale [8]. Golder et al. examine Twitter users' desire to follow another user connected by a path length of two. They examine the correlation between shared interests and reciprocated following on users' expressed inter-

[1] Following is a relatively passive activity and the establishment of a link between such users may misrepresent current attention to information in the network. Furthermore, follower networks typically do not account for the "unfriending" problem and the accumulation of dead links in a network can distort the representation of the true state of the system and spam.

[2] We observe approximately 35% of new links occurring between individuals connected by a path of length 2.

est to make a new link (i.e., follow) and suggest that mutuality (reciprocated attention) is correlated with increased desire to follow [39].

We organize our paper as follows: In Section 2, we describe our data, the sixteen similarity indices, and the evolutionary algorithm used for evolving the weights on these indices. In Section 3 we present our results and in Section 4 discuss the significance of these findings, as well as suggest future directions for further work in this area.

3.2 Methods

3.2.1 Data

Our data set consists of over 51 million tweets collected via the Twitter gardenhose API service from September 9, 2008 to December 1, 2008. This collection represents roughly 40% of all messages sent during this period (Table A1). Using the criteria defined by Bliss et al. [34], we construct reciprocal reply networks[3] as unweighted, undirected networks in which a link exists between nodes u and v if and only if these individuals exhibit reciprocal replies during the week under analysis (Fig. 3.1). These networks range in size from $N = 78296$ to $N = 155753$ nodes (Table A2).

Since our task is to predict links, we do not wish to confound our task with the problem of node appearance or removal. To this end, we find a core of 25,936 users who were active in each of networks in the period from September 9, 2008 to October 20, 2008 and a core of 44,439 users who were active in each of the weeks in the six weeks from October 21, 2008 and December 1, 2008. We train our link predictor on the new links that occur in a

[3]We also construct reply networks, whereby nodes represent users and directed, weighted links represent the number of replies sent from one individual to another during the week under analysis. Reply networks are used in the computation of the average path weight, one of our similarity indices.

given Week t (e.g., $e \in E_t \setminus E_{t-1}$) and validate on the new links that occur in week $t+1$ (e.g., $e \in E_{t+1} \setminus E_t$). We outline further details in the next two subsections.

Topological similarity indices (abbreviation)		
Jaccard Index (J)	$J(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\cup\Gamma(v)\rvert}$	Measures the probability that a neighbor of u or v is a neighbor of both u and v. This measurement is a way of characterizing shared content and has been shown to be meaningful in information retrieval [15].
Adamic-Adar Coefficient (A)	$A(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{log(\lvert\Gamma(z)\rvert)}$	Quantifies features shared by nodes u and v and weights rarer features more heavily [19]. Interpreting this in the context of neighborhoods, the Adamic-Adar Coefficient can be used to characterize neighborhood overlap between nodes u and v, weighting the overlap of smaller such neighborhoods more heavily.
Common neighbors (C)	$C(u,v) = \lvert\Gamma(u)\cap\Gamma(v)\rvert$	Measures the number of shared neighbors between u and v. Despite the simplicity of this index, Newman [9] documented that the probability of future links occurring in a collaboration network was positively correlated with the number of common neighbors.
Average Path Weight (P)	$P(u,v) = \frac{\sum_{p\in\mathcal{P}_2(u,v)\cup\mathcal{P}_3(u,v)} w_p}{\lvert\mathcal{P}_2(u,v)\rvert+\lvert\mathcal{P}_3(u,v)\rvert}$	Computes the sum of the minimum weights on the directed paths between u and v divided by the number of paths between u and v, where only paths of length 2 and 3 are considered due to the large size of this network. We take w_p to be the minimum weight of the edges in the path, in the spirit that a path's strength is only as strong as its weakest edge.
Katz (K)	$K = \sum_{n=1}^{\infty} \beta^n A^n$	Computed as such, the Katz is a global index [20]. This series converges to $(I-\beta A)^{-1} - I$, when $\beta < \max(\lambda(A))$. When $\beta \ll 1$ then K approximates the number of common neighbors. Due to the size of our network and computational expense of this index, we truncate to $n = 3$. We set $\beta = 1$ because we are not concerned with convergence & to emphasize the number of paths of length greater than two. Previous observations suggest that individuals who appear to be connected by a path length of n in Twitter RRNs may actually be connected by a path of shorter length due to role of missing data [34].
Preferential Attachment (Pr)	$Pr(u,v) = k_u \times k_v$	Gives higher scores to pairs of nodes for which one or both have high degree. This index arose from the observation that nodes in some networks acquire new links with a probability proportional to their degree [9] and preferential attachment random growth models [10].
Resource Allocation (R)	$R(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\lvert\Gamma(z)\rvert}$	Considers the amount of a given resource one node has and assumes that each node will distribute its resource equally among all neighbors [3].
Hub promoted Index (Hp)	$Hp(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\min\{k_u,k_v\}}$	First proposed to measure the topological overlap of pairs of substrates in metabolic networks, this index assigns higher scores to links adjacent to hubs since the denominator depends on the minimum degree of the two users [11].
Hub depressed Index (Hd)	$Hd(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\max\{k_u,k_v\}}$	When one of the nodes has large degree, the denominator will be larger and thus Hd is smaller in the case where one of the users is a hub [13].
Leicht-Holme-Newman Index (L)	$L(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{k_u k_v}$	Measures the number of common neighbors relative to the square of their geometric mean. This index gives high similarities to pairs of nodes that have many common neighbors compared to the expected number of such neighbors [14].

continued ...

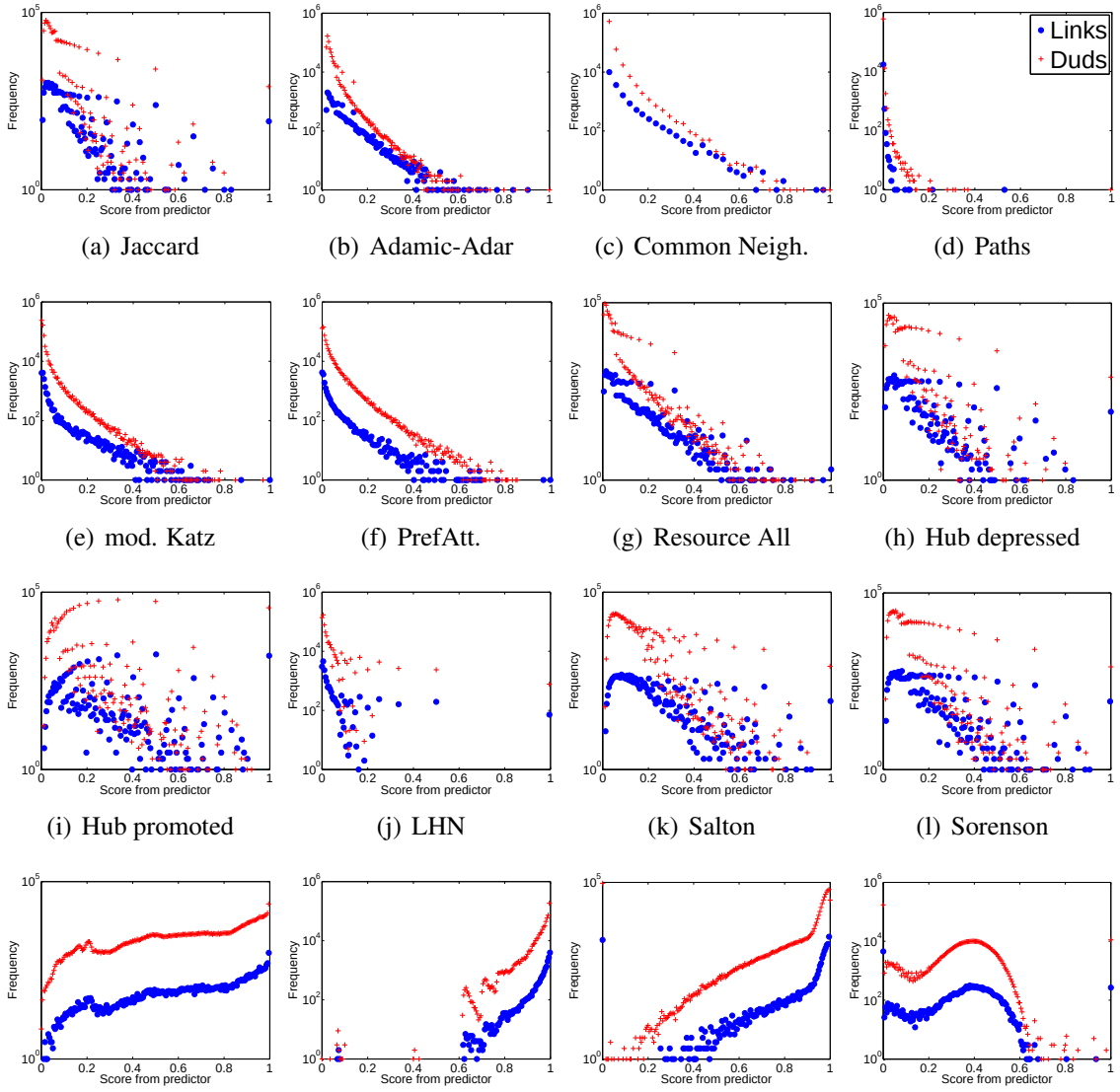
...continued

Topological similarity indices (abbreviation)		
Salton Index (Sa)	$Sa(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\sqrt{k_u k_v}}$	Measures the number of common neighbors relative to their geometric mean [15].
Sorenson Index (So)	$So(u,v) = \frac{2\lvert\Gamma(u)\cap\Gamma(v)\rvert}{k_u+k_v}$	Measures the number of common neighbors relative to their arithmetic mean. This index is similar to J, however J counts the number of (unique) nodes in the shared neighborhood. This index was previously used to establish equal amplitude groups in plant sociology based on the similarity of species [16].
Individual characteristics similarity indices		
Id similarity (I)	$I(u,v) = 1 - \frac{\lvert Id(u)-Id(v)\rvert}{\max\{\lvert Id(a)-Id(b)\rvert\}_{a,b\in V}}$	In 2008, user ids were numbered sequentially and a user's id served as a proxy for the relative length of time since opening a Twitter account. Id similarity characterizes the extent to which two individuals adopt Twitter simultaneously.
Tweet count similarity (T)	$T(u,v) = 1 - \frac{\lvert T(u)-T(v)\rvert}{\max\{\lvert T(a)-T(b)\rvert\}_{a,b\in V}}$	Tweet count $T(u)$ measures the number of Tweets we have gathered for node u in a given week. Tweet count similarity quantifies how similar two individuals' tweet counts are, with 1 representing identical tweet counts and 0 representing dissimilar tweet counts.
Happiness similarity (H)	$\mathrm{H}(u,v) = 1 - \frac{\lvert h(u)-h(v)\rvert}{\max\{\lvert h(a)-h(b)\rvert\}_{a,b\in V}}$	Building on previous work [40], happiness scores ($h(u)$ and $h(v)$) are computed as the average of happiness scores for words authored by users u and v during the week of analysis.
Word similarity (W)	$W(u,v) = 1 - \frac{1}{2}\sum_{n=1}^{50000}\lvert f_{u,n} - f_{v,n}\rvert$	From a corpus consisting of the 50,000 most commonly occurring words used in Twitter from 2008 through 2011 [40], the similarity of words used by u and v is computed by a modified Hamming distance, where $f_{u,n}$ represents the normalized frequency of word usage of the nth word by user u. The value of $W(u,v)$ ranges from 0 (dissimilar word usage) to 1 (similar word usage) [34].

Table 3.1: The sixteen similarity indices chosen for inclusion in the link predictor. We define the *neighborhood of node* u to be $\Gamma(u) = \{v \in V | e_{u,v} \in E\}$, where $G = (V, E)$ is a network, consisting of vertices (V) and edges (E). The degree of node u is represented by k_u, the adjacency matrix is denoted by A, and a path of length n between $u, v \in V$ is denoted as $\mathcal{P}_n(u, v)$.

3.2.2 Similarity indices

Similarity indices capture the shared characteristics or contexts of two nodes. We briefly describe 16 similarity indices chosen for inclusion in our link predictor, but wish to emphasize that any number of other similarity indices may be chosen for inclusion in the evolutionary algorithm. The choice of which similarity indices to include may largely depend on the metadata one has about the nodes and interactions, as well as the size of the network.



(a) Jaccard (b) Adamic-Adar (c) Common Neigh. (d) Paths

(e) mod. Katz (f) PrefAtt. (g) Resource All (h) Hub depressed

(i) Hub promoted (j) LHN (k) Salton (l) Sorenson

(m) Twitter Id similarity (n) Tweet count similarity (o) Happiness similarity (p) Word similarity

Figure 3.2: Similariy scores do not differentiate the link prediction signal. Scores for user-user pairs with path length two in Week 7, which exhibit a link (blue) and which did not (red) in Week 8. A higher score means that the user-user pair is more similar. For many indices, there are more "duds" than "links" for a given score. Indices for which there are "links" scoring higher than "duds" tend to exhibit a large, positive evolved coefficient (e.g., Adamic-Adar).

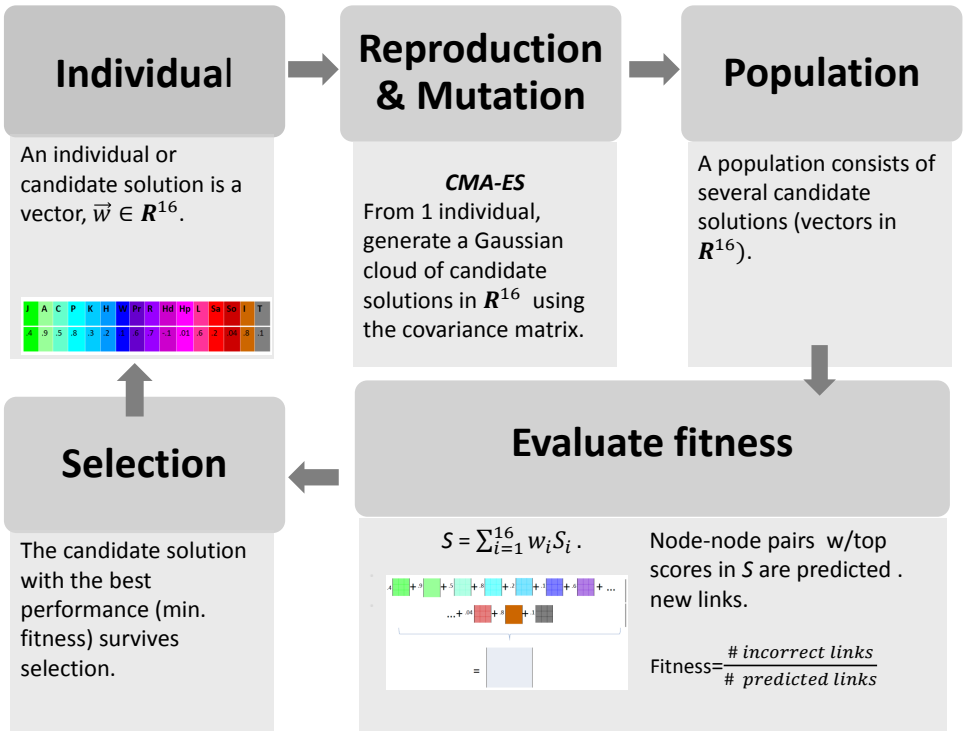


Figure 3.3: Link prediction with CMA-ES. An individual (or candidate solution) is a vector, $\vec{w} \in \mathbb{R}^n$, where n represents the number of indices used to constructor the predictor. We chose 16 such similarity indices. The initial individual is $\vec{w}_0$ where each entry is initialized between 0 and 1. From one individual, a Gaussian cloud of points in $\mathbb{R}^{16}$ is generated from the covariance matrix. This step mimics reproduction and mutation and creates a population of candidate solutions. Fitness is calculated for each candidate as the proportion of links incorrectly predicted, where a new link e_{ij} is predicted if s_{ij} is one of the top entries in matrix S. Selection occurs by taking the best candidate solution, $\vec{w} \in \mathbb{R}^{16}$. This one individual survives the generation and the cycle is repeated.

Topological similarity indices may be characterized by local, quasi-local, or global measures. Since global similarity measures (i.e., Katz, SimRank, and Matrix Forest Index) are computationally laborious for large networks [13], we forgo these measures in lieu of local topological indices. For node similarity we calculate four indices: Twitter Id similarity, tweet count similarity, word similarity and happiness similarity. All of these indices are described in Table 3.1. We then rescale the computed scores to range from 0 to 1, inclusive, and store as $N \times N$ sparse matrices, hereafter referred to as S_i, for $i = 1, 2, \ldots, 16$.

We depict frequency plots for the computed similarity indices in Figure 3.2. These plots demonstrate that none of the similarity indices separate the newly formed "links" (user-user pairs who are separated by a minimal path of length 2 at t and a path of length 1 at $t + 1$) and "duds" (user-user pairs who are separated by a minimal path of length 2 at t and a path of length $\delta \neq 1$ at $t + 1$). This lack of separation is one indication that a predictor which combines information from several indices may improve link prediction efforts. Figure 3.2 also reveals that the manner in which the predictors should be combined is not as straightforward as one might envision. For example, some similarity indices, such as Adamic-Adar (Fig. 3.2b) and Resource Allocation (Fig. 3.2i) show potential for differentiating links and duds. Other indices, such as Twitter Id similarity (Fig. 3.2o) maintain a greater number of duds than links, across all scores. This is a result of the large class imbalance between the number of potential user-user pairs for new links and the actual numbers of new links formed, a common occurrence in large, sparse networks.

3.2.3 Evolutionary algorithm

Evolutionary algorithms take inspiration from biological systems whereby individuals representing candidate solutions evolve over generational time via selection, reproduction,

mutation, and recombination (Fig. 3.3). In our task, we construct a linear combination of similarity indices, S_i, and use an evolutionary strategy to evolve the coefficients, w_i, used in computing a score matrix, S,

$$S = \sum_{i=1}^{16} w_i S_i, \tag{3.1}$$

for which the minimum error in link prediction is desired.

Our task is essentially an optimization problem. Our choice for CMA-ES stems from its efficiency in finding real valued solutions in noisy landscapes [41]. In contrast to gradient descent approaches for finding optimal solutions, CMA-ES is not reliant on assumptions of differentiability nor continuity of the fitness landscape. Our method requires no heuristics, which is an advantage over many existing supervised learning methods (e.g., SVM) that require extensive parameter tuning and kernel selection [29]. Additionally, our method is flexible and allows for any similarity index to be substituted into or added to the evolutionary algorithm. Ideally, the transparency of the evolved "best" predictors will help illustrate possible driving mechanisms behind the network's evolution. This method is also one of the best evolutionary algorithms for finding optima of real valued solutions due to its fast convergence.[4] We refer the interested reader to [42] for more detail regarding the CMA-ES algorithm.

Figure 3.3 outlines our implementation of CMA-ES for link prediction. Before employing the evolutionary algorithm, all similarity indices are computed and stored as $N \times N$ sparse matrices, S_i for $i = 1, 2, \ldots, 16$. The evolutionary algorithm begins with a candidate solution termed an "individual" in the language of evolutionary computation. Entries of $\vec{w}$ are initially set to real values between 0 and 1 chosen from a uniform random distribution.

[4]Here, we refer to fast convergence in generational time. The CPU time for one generation of our CMA-ES implementation for link prediction was 13 seconds.

These values are not constrained during evolution. Using CMA-ES with both rank-1 and rank-μ updates[5] we evolve $\vec{w} = \langle w_1, w_2, \ldots, w_{16} \rangle \in \mathbb{R}^{16}$ over 250 generations [29]. At each generation, a population of candidate solutions is selected from a multivariate Gaussian cloud[6] surrounding the "individual" surviving the previous generation.

Each candidate solution in the "population" is assessed for fitness and the individual with the best fitness survives the generation. The standard implementation of CMA-ES selects the "best solution" as that which minimizes fitness. As such, our fitness function[7] computes the link prediction error for each $\vec{w} \in \mathbb{R}^{16}$. One of the difficulties with CMA-ES is the potential to be trapped in local optima. To avoid this, we perform 100 restarts, a technique suggested by Auger and Hansen [43].

3.2.4 Cross referencing links

From the 100 best solutions evolved via CMA-ES for each of the four fitness functions (e.g., where the top 20, 200, 2000 or 20000 scores are used to predict future links) we cross-reference the top N scoring user-user pairs. The user-user pairs which are most heavily cross-referenced (i.e., links which most models agree upon) are those for which we predict a link. In addition to the 400 best evolved predictors, we also feed in information from the Resource Allocation similarity index when prediction top N <10 because of the high performance of this index for predicting the top 10 or fewer links on training sets.

[5]Briefly, rank-1 updates utilize information about correlations between generations, which is helpful for evolution with small populations of candidate solutions. Rank-μ updates utilize information from the current generation, which helps speed up the algorithm for large populations.

[6]We use the default population size of $4 + \lfloor 3 \log(m) \rfloor$, for solutions in $\mathbb{R}^m$, from Hansen's source code available at `https://www.lri.fr/~hansen/cmaes_inmatlab.html` (last accessed on October 1, 2012). Increasing the population size did not improve our results.

[7]for each of four fitness functions fitness_{20}, fitness_{200}, fitness_{2000}, fitness_{20000} where the subscript denotes the top N scoring user-user pairs (e.g., predicted links). By incorporating fitness functions which operate at different scales, we investigate the sensitivity of the top N on the link predictor's performance in validation.

3.3 Results

Our overall finding is that the evolved predictor consisting of all sixteen similarity indices outperformed all other combined and individual indices on the training data when training occurred on a given week's RRN. In Figure 3.4, we present the results for fitness_{20} during training on new links formed from Week 7 to Week 8. The solid black curve depicting the

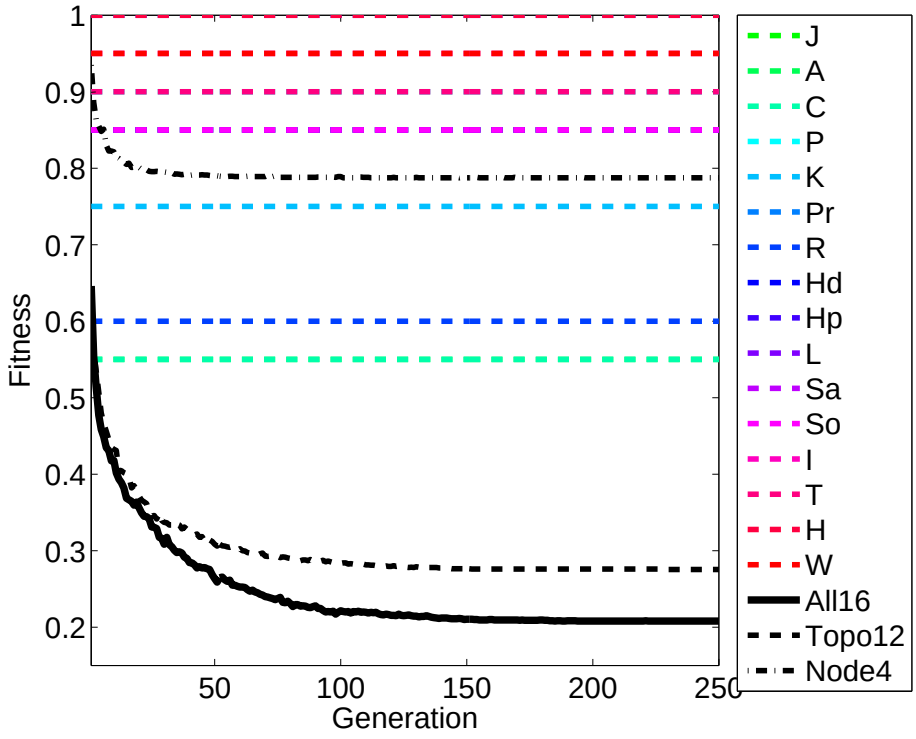
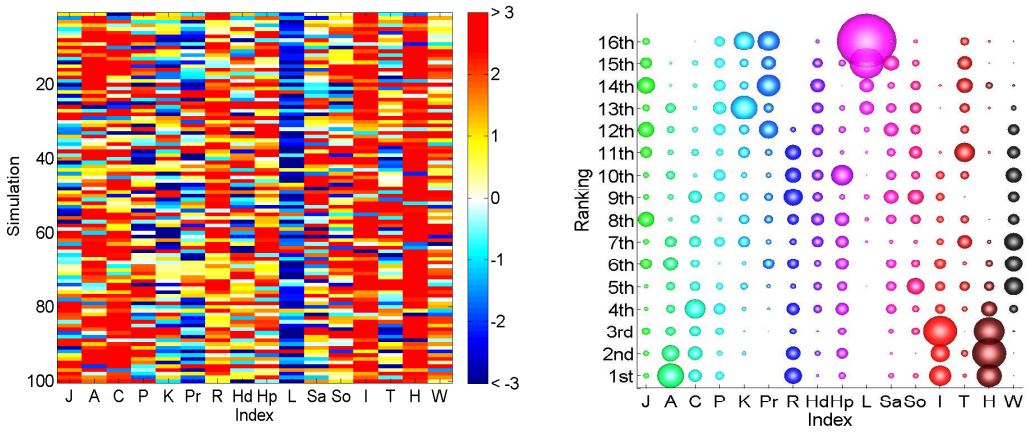


Figure 3.4: Mean best fitness computed from 100 simulations of CMA-ES. The algorithm trains on the new links that occur in Week 8 (i.e., links present in Week 8 that were not present in Week 7) using fitness_{20}. The evolutionary algorithm seeks to minimize fitness (i.e., minimize the proportion of falsely predicted links). We compare each individual index (shown in color), along with the three evolved predictors (shown in black): "all16" (all 16 indices), "topo12" (12 topological indices), and "node4" (4 individual similarity indices). The "all16" predictor performs the best, followed by the "topo12" predictor.

"all16" predictor shows that while the average fitness at generation 1 for the 100 candidates

was far worse ($\approx$ 0.65) than several similarity indices such as Adamic-Adar ($\approx$ 0.55), Common neighbors ($\approx$ 0.55) and Resource Allocation $\approx$ 0.60), convergence to a far better set of solutions occurred within 100 generations ($\approx$ .22). The combination of the twelve topological indices outperformed all individual indices, but was outperformed by the all16 predictor. This difference is most pronounced for the top N=20 cases, however this trend holds true for the other fitness functions (Appendix, Fig. 3.A1).



(a) 100 evolved best "inidividuals" from CMA-ES

(b) Frequency plot for ranked coefficients, w_i corresponding to similarity indices

Figure 3.5: Characterizing the 100 best "individuals" from CMA-ES. (a.) Presentation of the best solutions evolved from each of 100 simulations using fitness$_{20}$ and the "all16" predictors to predict new links that occurred from Week 7 to 8. (b.) Frequency plot of ranked coefficients from (a.), where 1st place represents large, positive coefficients and 16th place represents large, negative coefficients. Disk size indicates the fraction of times an index received a given ranking. Adamic-Adar, Happiness similarity, Resource Allocation and Twitter Id similarity were the most commonly occurring indices ranked 1st (largest, positive) coefficient, and LHN often evolved to the largest, negative coefficient. This suggests possible mechanisms which may have been driving the evolution of the network during this time period. J=Jaccard, A=Adamic-Adar, C=Common neighbors, P=Paths, K=Katz, Pr=Preferential attachment, R=Resource allocation, Hd=Hub depressed, Hp=Hub promoted, L=Leicht-Holme-Newman, Sa=Salton, So=Sorenson, I=Twitter id similarity, T=Tweet count similarity, H=Happiness similarity, W=Word similarity.

Our interest extends beyond an analysis of the proportion of links correctly predicted. We reveal the constituents of our link predictor ($\vec{w} \in \mathbb{R}^{16}$) as a means to gain an (initial) understanding of the mechanisms which may be driving the evolution of Twitter RRNs. In this spirit, we present two visualizations which capture this information. For illustration purposes, we highlight the results from Week 8, using a fitness function which selects the top 20 scores as new links, in Figure 3.5.

Figure 3.5a shows all 100 solutions which evolved after 250 generations of CMA-ES, $\vec{w}$, as horizontal rows. The ith column signifies the w_i coefficient used in the linear combination of the weights. The color axis reveals the value of ith coefficient. Several trends are worth noting here. First, there is considerable variability between the 100 evolved best candidates. Second, despite this variability, Adamic-Adar, Common neighbors, Resource Allocation, Happiness, and Twitter Id similarity columns have many more positive values than negative. On the other hand, the coefficient for the Leicht-Holme-Newman index often evolved to a large negative weight. This signifies that user-user pairs which had high scores for the indices which evolved large, positive weights (e.g., Adamic-Adar, Common neighbors, Resource Allocation, Happiness, and Id similarity) and low scores for the indices which evolve large, negative weights (e.g., Leicht-Holme-Newman) were more likely to exhibit a future link.

We also visualize the relative ranking of the indices by their coefficients the Fig. 3.5b (and corresponding plots in the Appendices 3.A2–3.A5). Ordering the coefficients from greatest (most positive in 1st place) to least (most negative in 16th place) reveals that Adamic-Adar, Common neighbors, Resource Allocation, Happiness, and Twitter Id similarity often occupied the 1st-4th rankings (i.e., indices with the largest positive contribution, whereas LHN was often in 16th place (the largest negative weight). Other indices showed

considerable variability in their ranking. We explore the implications of these findings in our discussion.

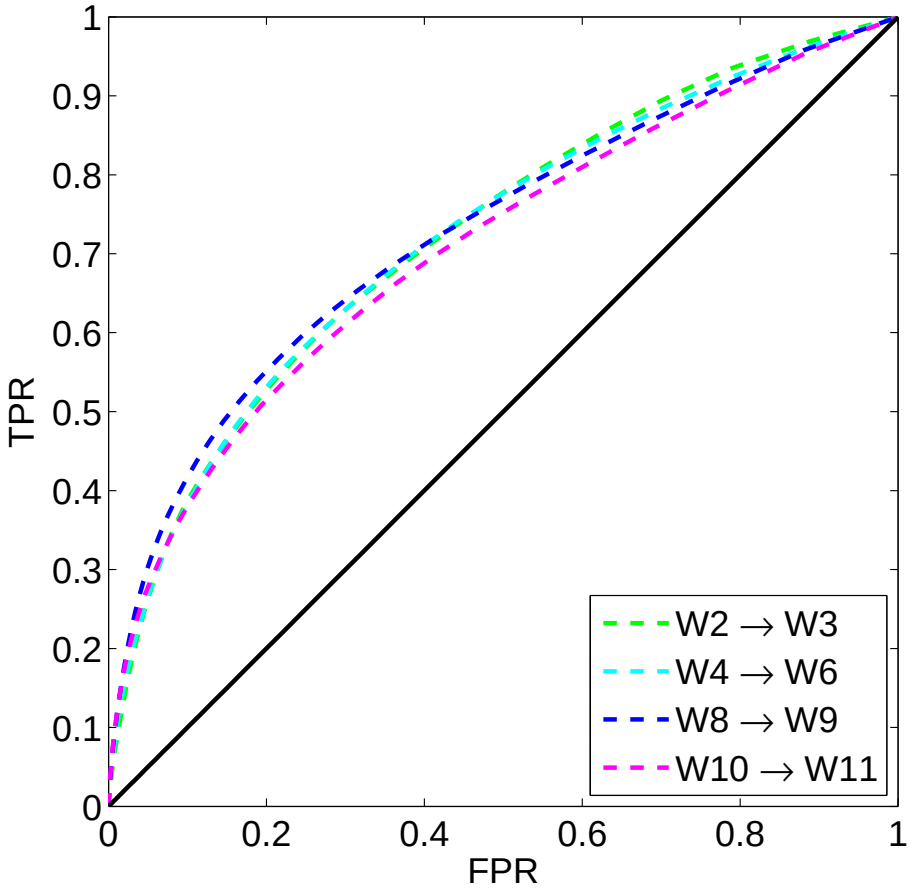


Figure 3.6: Receiver Operating Curve (ROC) for the all16 predictor using fitness_{20000}. $AUC_{\text{Week 2}\longmapsto 3} = .723$, $AUC_{\text{Week 4}\longmapsto 5} = .721$, $AUC_{\text{Week 8}\longmapsto 9} = .726$,, and $AUC_{\text{Week 10}\longmapsto 11} = .707$.

The ROC curve demonstrates that the true positive rate is considerably larger than the false positive rate ($TPR > FPR$) (Fig. 3.6). We find AUC scores greater than 0.7 for all weeks in the validation set, suggesting that our predictor performs quite well, especially compared to other work with Twitter follower networks which did not suffer from missing data issues [23]. We discuss these implications further in Section 4.

For large, sparse networks, the negative class is often much larger than the positive class. In our case, the number of new links (positive class) is on the order of 10^4, whereas the number of potential links which do not exhibit future links (negative class) is on the order of 10^8. Given this imbalance, measures such as accuracy, negative predictive value, and specificity will be very close to 1, even for random link predictors. As suggested by Wang et al. [4], more emphasis should be placed on recall and precision due to the large class imbalance between positives and negatives. The tunable parameter β allows for unequal weighting on recall vs. precision:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}. \tag{3.2}$$

In some applications, false positives ("false alarms") may be relatively costless, whereas false negatives ("misses") may pose an imminent threat. In these cases, recall is much more important than precision and setting $\beta > 1$ will weight recall more heavily in the F_β score. In contrast, other applications may involve scenarios where false positives are costly to explore and a small number of links, for which we are fairly certainly about, is highly prized. In these cases, one can set $\beta < 1$ to place more importance on precision.

Tuning β to one of 0.5, 1 or 2, we find that the F_1 peaks around top $N \approx 10^4$ (Fig. 3.7). F-scores are higher for weeks during which we received a higher percentage of tweets from the Twitter API service. For example, $F_{0.5} = 0.203, F_1 = .177, F_2 = .142$, and $F_{0.5} = 0.226, F_1 = .181, F_2 = .143$ for links which occurred from Weeks 8 to 9 and Weeks 10 to 11, respectively. In Week 5, we received a far smaller percentage of tweets. F-scores for new links occurring from Weeks 4 to 5 are $F_{0.5} = 0.184, F_1 = .152, F_2 = .128$.

Figure 3.8 depicts the precision of the predicted links as a function of the top N scoring user-user pairs. High precision is achieved for the fitness function which operates by

selecting the top 20 scoring user-user pairs, which is often the region of interest. Precision is lower for predicted links from Week 4 to 5, a week in which we received a very low percentage of tweets from the Twitter API service, and higher for predicted links from Week 8 to 9 and Week 10 to 11, weeks for which we received a higher percentage of tweets from the Twitter API service (see Table A1). We also compute negative predictive value, and find this is consistently close to 1 due to the large true negative class. Specificity and accuracy are close to 1 for nearly all values of top N links predicted, except for particularly large N ($> 10^4$). This is due to the large class imbalance of true negatives (TN), which dominate the numerator and denominator of these calculations.

3.3.1 Exploring the impact of missing data

During the twelve week period from September 9, 2008 - Dec 1, 2008 we received approximately 40% of all tweets from Twitter's API service (Table A1). There are therefore both individuals and interactions that are unaccounted for in our training and validation period. Consequently, there are individuals who are connected by a path of length two in the true network, but which appear to be connected by a longer path because we have not captured interactions for intermediaries.

We explore the potential impact of missing tweets on our predictor by randomly selecting 50% of our observed tweets and constructing the reciprocal reply subnetworks for Weeks 1 through 12. The evolutionary algorithm trains and validates on these subnetworks. For clarity, we denote G for our observed networks and G^s, for our subnetworks. We identify the percent of links which are labeled as false positives in G^s and true positive in G. This occurs precisely because our link predictor suggested a link which was actually correct, but for which an incomplete data set caused the link to be classified as a false positive.

As such, we are underestimating the success of our link prediction method. Given a more complete data set, our results would most likely be better than we report here.

We next investigate the effects of missing data on our predictor, under the condition that 50% of the Tweets have been removed. We observe that the number of correctly predicted links is hindered by the missing data, and the proportion of links which are incorrectly termed "false-positive" because they are actually links in the weekly network containing a more complete data set is roughly 10% (Fig. 3.9). This result from bootstrapping suggests that the performance of our predictors is a lower bound on performance, i.e., true precision and recall are most likely better than we report.

3.3.2 Comparison to other methods

Other studies in the area of link prediction have reported the factor improvement over random link prediction [1, 4]. We follow suit and compute the factor improvement of our predictor over a randomly chosen pair of users. The probability that a randomly chosen pair of individuals who are not connected in week i become connected in week $i+1$ is $\frac{|\text{Edges}_{\text{new}}|}{\binom{|V(G)|}{2}-|\text{Edges}_{\text{old}}|}$. There are 44,439 nodes in the validation set and, as a sample calculation, 71,927 edges in week 7. There are 53,722 new links that occur from Week 7 to 8. Thus, the probability of a randomly chosen pair of nodes from Week 7 exhibiting a link in Week 8 is approximately $\frac{53{,}722}{\binom{44{,}439}{2}-71{,}927} \approx .0054\%$.

We observe significant factors of improvement over randomly selected new links, usually on the order of 10^4 for top N <20 (Fig. 3.10). We notice that Resource Allocation outperformed other similarity indices when used in isolation to select the top 5 links during training and have included this in the cross-validation (Predictor$_{RA}$) step for selecting the top 10 (or fewer) links. We observe that the combined predictor outperforms indices used in

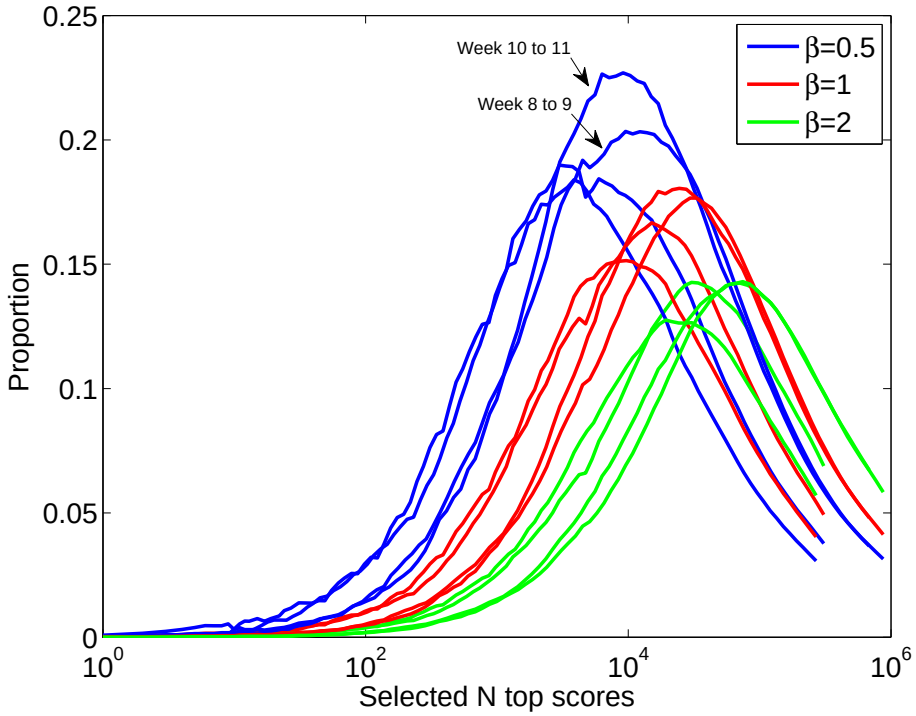


Figure 3.7: F_β scores for each of the validation sets. When $\beta = 1$, precision and recall are weighted equally. $\beta > 1$ weights recall ($TPR = \frac{TP}{TP+TP+FN}$), whereas $\beta < 1$ places more importance on precision ($PPV = \frac{TP}{TP+FP}$). Our predictor performs better with respect to precision and peaks for values on the order of 10^3. The standard F_1 score peaks around 10^4 and compares favorably with the work of [23]. The highest F_β scores are found for $W10 \rightarrow 11$.

isolation most choices of top N link prediction. Due to the recent interest in using network flow measures, we also compare our predictors to *propflow* restricted to a path of length two, a method proposed by Lichtenwalter et al. [17]. Our method strongly outperforms this index.

Lastly, we compare our results to those obtained by training a binary decision tree classifier.[8] Typically, balanced classes are used in training binary decision trees in order

[8]We use Matlab's implementation of binary classification trees to train on new links that form from Week 7 to Week 8.

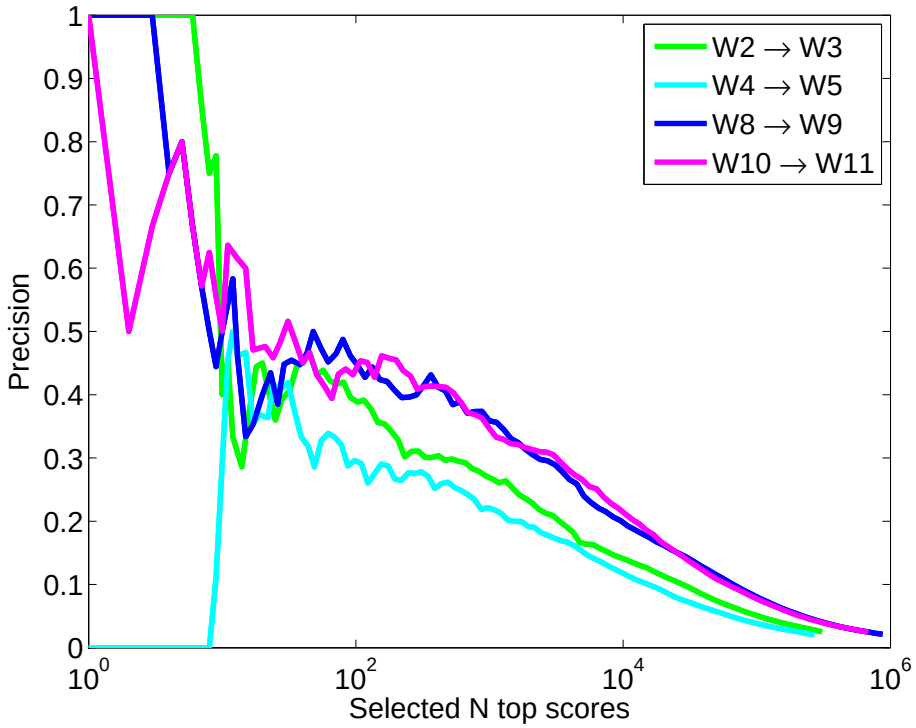


Figure 3.8: Precision for the predicted links in the validation sets. High precision is achieved for top$N$ < 20, which is often the region of interest. The precision for predicted links in $W4 \rightarrow W5'$ is lower than the other weeks and this may be due to missing data for those weeks (see Table A2).

to overcome problems associated with unbalanced classes [17, 44, 45]. We note that since our method for link prediction operates on all node-node pairs separated by path length two (e.g., highly unbalanced classes), we train our binary decision tree on unbalanced classes to avoid confounding our comparison with issues related to balanced and unbalanced classes. Furthermore, we set our method to select the top$N$=7417 links, which provides for roughly the same number of true positives as identified by the binary decision tree classifier. Table 3.2 reveals the results of this comparison. With this choice of top$N$, our approach performs slightly better across several indicators, such as accuracy and recall. Most notably, our precision is nearly three times as great as that obtained from our binary decision

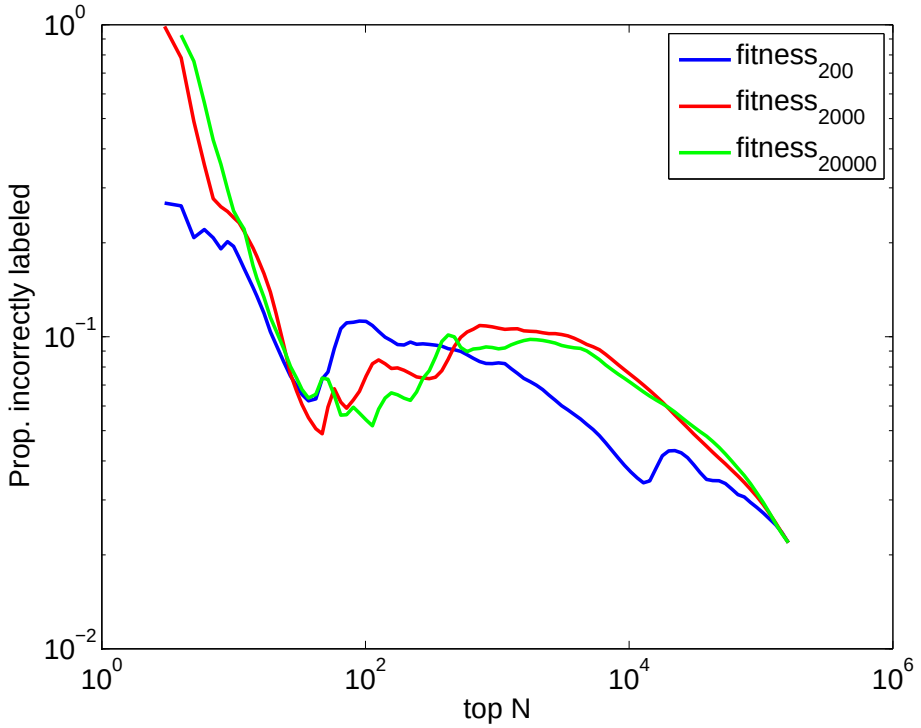
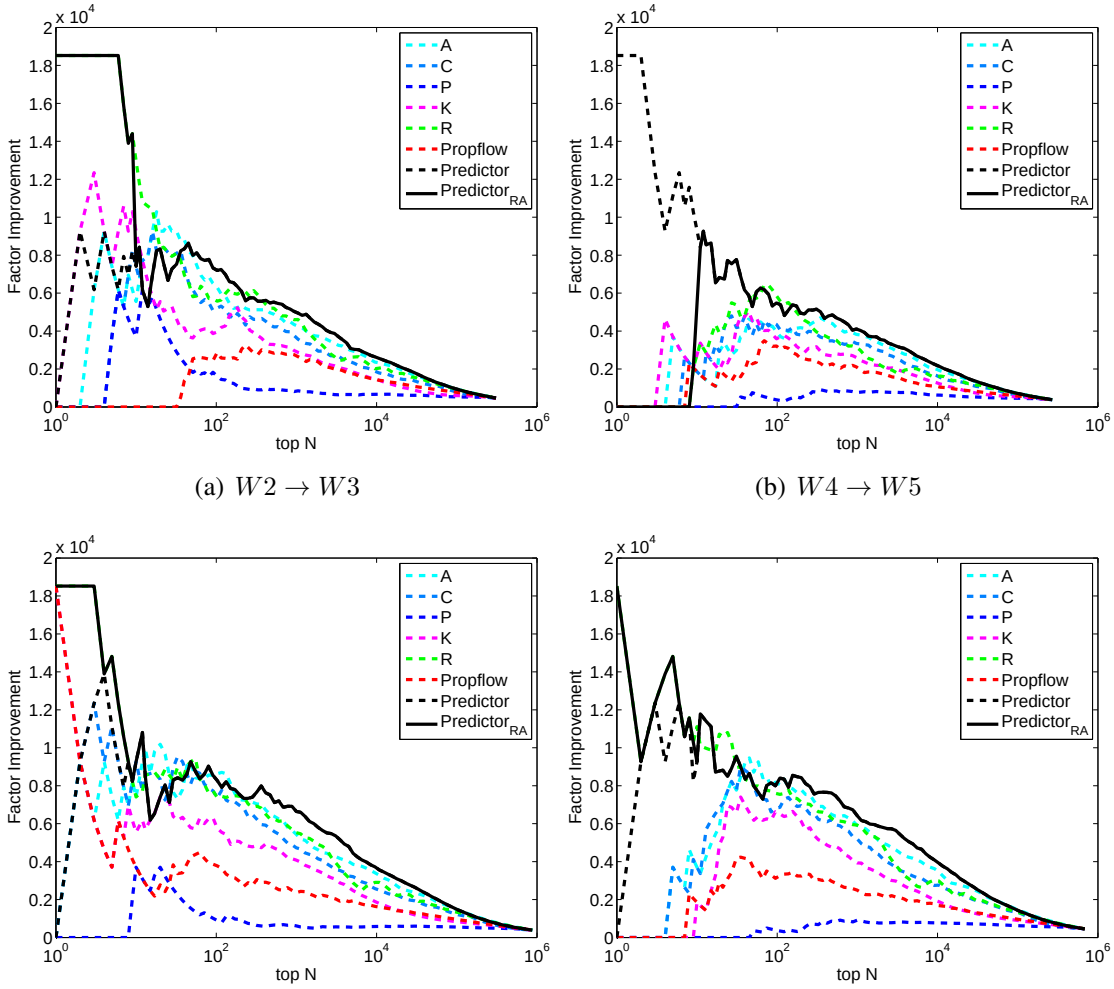


Figure 3.9: Proportion of incorrectly labeled false positives due to incomplete data. To test the effect of missing data, we remove 50% of our observed tweets and recreate networks using this subsample of the data for Week 7 to 8.

tree. Our false discovery rate is lower than that obtained for binary decision trees and this may be simply due to our taking a top$N$ approach to link prediction, which inherently limits the number of false positives by tuning the top$N$ links to predict. We discuss these results in more detail in the next section.

3.4 Discussion

Our measures perform quite well in comparison to other researchers working in the area of link prediction for Twitter. Rowe, Stankovic, and Alani [23] explore topological and individual specific similarity indices (words and topic similarity) in an effort to predict



(a) $W2 \rightarrow W3$ (b) $W4 \rightarrow W5$

(c) $W8 \rightarrow W9$ (d) $W10 \rightarrow W11$

Figure 3.10: Factor improvement over randomly selected user-user pairs. Large factor improvements are exhibited for predicting the top N links, with notable peaks for N <100. The combined predictor outperforms the Common neighbors, Adamic-Adar, Paths, Katz, and Resource Allocation indices used in isolation over most choices for the top N links predicted.

following behavior. They find an $AUC < 0.6$ whereas we find $AUC > 0.7$ for all experiments. Yin, Hong, and Davison [8] develop a structure based link prediction model and report F-scores on the order of $F = .190$ for Twitter follower networks. These networks do not suffer from incomplete data in the same way that Twitter reciprocal reply networks

	Binary Decision Tree	CMA-ES
Accuracy	0.9555	0.9741
Precision	0.0894	0.2131
Recall (true pos. rate)	0.0694	0.0858
False positive rate	0.0197	0.0068
False discovery rate	0.9106	0.7869

Table 3.2: Comparison of binary decision trees vs. CMA-ES for top$N$ link prediction. CMA-ES (with top$N$=7417) slightly outperforms binary decision trees trained on new links that form from Week 7 to Week 8. We note that unbalanced classes are used in both cases.

do. Our predictor performs comparatively well, with scores ranging from F_1 =0.152 for validation on new links occurring from Week 4 to 5, a week for which we obtained approximately 24% of all tweets, to F_1 =0.181 for validation on new links occurring from Week 10 to 11, a week for which we obtained approximately 48% of all tweets.

We have developed a meaningful link predictor for Twitter reciprocal reply networks, a social subnetwork consisting of individuals who demonstrate active and ongoing engagement. We were able to achieve a factor of improvement over random link selection on the order of 10^4 for the top 20 (or fewer) links predicted and 10^3 over several orders of magnitude for the top N links predicted.

Wang et al. [4] examine a social network constructed from mobile phone call data and find a factor improvement of approximately 1.5×10^3. To compare our work, however, one must standardize for the number of nodes in the network.[9] Upon doing so, we find our factor improvement is an order of magnitude higher.

We compare our results to other approaches, such as *propflow* and binary decision trees. As suggested by others and observed here, link prediction in large, sparse networks suffers from problems related to unbalanced classes. As such, we caution the interpretation of our

[9]These researchers report 579,087,610 potential new links and a factor improvement of 1500. Rescaling the factor improvement for networks of the same size amounts to computing the probability of a randomly predicted link being correct.

results in comparision to industry standards, such as binary decision trees. Future work may improve upon our methods by using balanced classes in the evolution of coefficients over generational time in CMA-ES. Incorporating these strategies and others may allow for more insightful comparisons between our methods and other supervised learning approaches.

One of the most intriguing aspects of this work is the detection of similarity indices which evolve to have large, positive weights in our link predictors. Perhaps the most notable similarity index for which this is the case is the Resource Allocation index. Resource allocation considers the amount of resource one node has and assumes that each node will distribute its resource equally among all neighbors [3]. Considering the limits to time and attention an individual has, this may be suggestive of a mechanism by which users limit their interaction, a result suggested by Gonçalves et al. [46] and also noted by [34] in Twitter RRNs.

In addition to suggesting that our work is comparable to or an improvement upon other work which combines measures via supervised learning, we present a method which is transparent and transferable. Future work may involve the inclusion of geospatial data [47] or community structure to predict links. Efforts to consider the persistence or decay of links over time, or inconsistencies in flow rates [48] could also prove fruitful.

3.5 Acknowledgments

The authors acknowledge the Vermont Advanced Computing Core which is supported by NASA (NNX-08AO96G) at the University of Vermont for providing High Performance Computing resources that have contributed to the research results reported within this paper. CAB and PSD were funded by an NSF CAREER Award to PSD (# 0846668). CMD,

PSD, and MRF were funded by a grant from the MITRE Corporation. The authors thank Brian Tivnan and Maggie J. Eppstein for their helpful suggestions.

3.6 References

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

[2] Z. Lu, B. Savas, W. Tang, and I.S. Dhillon. Supervised link prediction using multiple sources. In *2010 IEEE 10th International Conference on Data Mining (ICDM)*, pages 923–928. IEEE, 2010.

[3] T. Zhou, L. Lü, and Y.C. Zhang. Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(4):623–630, 2009.

[4] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011.

[5] I. Esslimani, A. Brun, and A. Boyer. Densifying a behavioral recommender system by social networks link prediction methods. *Social Network Analysis and Mining*, 1(3):159–172, 2011.

[6] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 635–644. ACM, 2011.

[7] V. Leroy, B. B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 393–402. ACM, 2010.

[8] D. Yin, L. Hong, and B. D. Davison. Structural link analysis and prediction in microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1163–1168, New York, NY, USA, 2011. ACM.

[9] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64:025102, Jul 2001.

[10] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614, 2002.

[11] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[12] J. Wang and L. Rong. Similarity index based on the information of neighbor nodes for link prediction of complex network. *Modern Physics Letters B*, 27(06), 2013.

[13] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[14] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, volume 1, pages 296–304. San Francisco, 1998.

[15] G. Salton and M. J. McGill. In *Introduction to modern information retrieval*. McGraw-Hill, Inc., 1986.

[16] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. skr.*, 5:1–34, 1948.

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

[18] Y. Yang, N. V. Chawla, Y. Sun, and J. Han. Predicting links in multi-relational and heterogeneous networks. In *Proceedings of the 12th IEEE International Conference on Data Mining*, ICDM 12, pages 755–764, 2012.

[19] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211 – 230, 2003.

[20] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

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

[22] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6(2):9:1–9:33, 2012.

[23] M. Rowe, M. Stankovic, and H. Alani. Who will follow whom? Exploiting semantics for link prediction in attention-information networks. In *Proceedings of the 11th*

International Conference on The Semantic Web - Volume Part I, ISWC'12, pages 476–491, 2012.

[24] C. J. Hutto, S. Yardi, and E. Gilbert. A longitudinal study of follow predictors on Twitter. In *CHI 2013 "Changing Perspectives," First ACM European Computing Research Congress*, 2013.

[25] Daniel M Romero, Chenhao Tan, and Johan Ugander. On the interplay between social and topical structure. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM, 2013.

[26] Z. Yin, M. Gupta, T. Weninger, and J. Han. LINKREC: a unified framework for link recommendation with user attributes and graph structure. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1211–1212. ACM, 2010.

[27] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM06: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[28] C. J. C Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[29] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[30] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in Twitter: The million follower fallacy. 2010.

[31] 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*, WWW '10, pages 591–600, New York, NY, USA, 2010.

[32] B. H. Huberman, D. H. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *CoRR*, abs/0812.1045, 2008.

[33] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[34] C. A. Bliss, I. M. Kloumann, K. D. Harris, C. M. Danforth, and P. S. Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, 3(5):388 – 397, 2012.

[35] A. Rapoport. Mathematical models of social interaction. *Handbook of Mathematical Psychology*, 2:493–579, 1963.

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

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

[38] D. M. Romero and J. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 138–145, 2010.

[39] S. A. Golder and S. Yardi. Structural predictors of tie formation in Twitter: Transitivity and mutuality. In *Proceedings of the 2010 IEEE Second International Con-*

ference on Social Computing, SOCIALCOM '10, pages 88–95, Washington, DC, USA, 2010. IEEE Computer Society.

[40] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS one*, 6(12):e26752, 2011.

[41] T. Suttorp, N. Hansen, and C. Igel. Efficient covariance matrix update for variable metric evolution strategies. *Machine Learning*, 75(2):167–197, 2009.

[42] Nikolaus Hansen. The CMA evolution strategy: A tutorial. *Vu le*, 29, 2005.

[43] A. Auger and N. Hansen. A restart CMA evolution strategy with increasing population size. In *IEEE Congress on Evolutionary Computation*, volume 2, pages 1769–1776. IEEE, 2005.

[44] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

[45] D. A. Cieslak and N. V. Chawla. Learning decision trees for unbalanced data. In *Machine Learning and Knowledge Discovery in Databases*, pages 241–256. Springer, 2008.

[46] B. Gonçalves, N. Perra, and A. Vespignani. Modeling users' activity on Twitter networks: Validation of Dunbar's Number. *PLoS one*, 6, 08 2011.

[47] M. R. Frank, L. Mitchell, P. S. Dodds, and C. M. Danforth. Happiness and the patterns of life: A study of geolocated tweets. *Nature Scientific Reports*, 3, 2013.

[48] J. P. Bagrow, S. Desu, M. R. Frank, N. Manukyan, L. Mitchell, A. Reagan, E. E. Bloedorn, L. B. Booker, L. K. Branting, M. J. Smith, B. F. Tivnan, C. M. Danforth,

P. S. Dodds, and J. C. Bongard. Shadow networks: Discovering hidden nodes with models of information flow. *arXiv preprint, arXiv:1312.6122*, 2013.

3.7 Appendix

Week	Start date	# Obsvd. Msgs. $\times 10^6$	# Total Msgs. $\times 10^6$	% Obsvd.	# Replies $\times 10^6$	% Replies
1	09.09.08	3.14	7.26	43.2	0.88	28.1
2	09.16.08	3.36	8.31	40.4	0.90	26.9
3	09.23.08	3.43	8.89	38.6	0.90	26.2
4	09.30.08	3.33	9.06	36.8	0.89	26.6
5	10.07.08	2.33	9.38	24.8	0.64	27.5
6	10.14.08	4.39	9.87	44.4	1.24	28.3
7	10.21.08	4.70	10.01	47.0	1.35	28.8
8	10.28.08	5.74	10.34	55.5	1.64	28.5
9	11.04.08	5.58	11.14	50.1	1.63	29.3
10	11.11.08	4.70	9.88	47.6	1.42	30.2
11	11.18.08	5.48	11.34	48.3	1.67	30.5
12	11.25.08	5.71	11.47	49.8	1.73	30.2

Table 3.A1: Number of "observed" messages in our database. The number of "observed" messages in our database comprise a fraction of the total number of Twitter message made during period of this study (September 2008 through November 2009). While our feed from the Twitter API remains fairly constant, the total # of tweets grows, thus reducing the % of all tweets observed in our database. We calculate the total # of messages as the difference between the last message id and the first message id that we observe for a given month. This provides a reasonable estimation of the number of tweets made per month as message ids were assigned (by Twitter) sequentially during the time period of this study. We also report the number observed messages that are replies to specific messages and the percentage of our observed messages which constitute replies.

Week	Start date	N	$< k >$	$k_{\max}$	C_G	Assort	# Comp.	S
1	09.09.08	95647	2.99	261	0.10	0.24	10364	0.71
2	09.16.08	99236	2.95	313	0.10	0.24	11062	0.71
3	09.23.08	99694	2.90	369	0.09	0.13	11457	0.70
4	09.30.08	100228	2.87	338	0.09	0.13	11752	0.69
5	10.07.08	78296	2.60	241	0.09	0.21	11140	0.63
6	10.14.08	122644	3.20	394	0.09	0.14	12221	0.74
7	10.21.08	130027	3.30	559	0.08	0.09	12420	0.75
8	10.28.08	144036	3.56	492	0.08	0.14	12319	0.78
9	11.04.08	145346	3.54	330	0.08	0.19	12597	0.78
10	11.11.08	136534	3.35	441	0.08	0.12	12972	0.76
11	11.18.08	153486	3.46	444	0.08	0.13	13594	0.77
12	11.25.08	155753	3.46	1244	0.06	0.00	14122	0.77

Table 3.A2: Network statistics for reciprocal-reply networks by week. As Twitter popularity grows, so does the number of users (N) in the observed reciprocal-reply network. The average degree ($< k >$), degree assortativity, the number of nodes in the giant component (# Comp.), and the proportion of nodes in the giant component (S) remain fairly constant, whereas the maximum degree ($k_{\max}$) shows a great deal of variability from month to month. Clustering (C_G) shows a slight decrease over the course of this period.

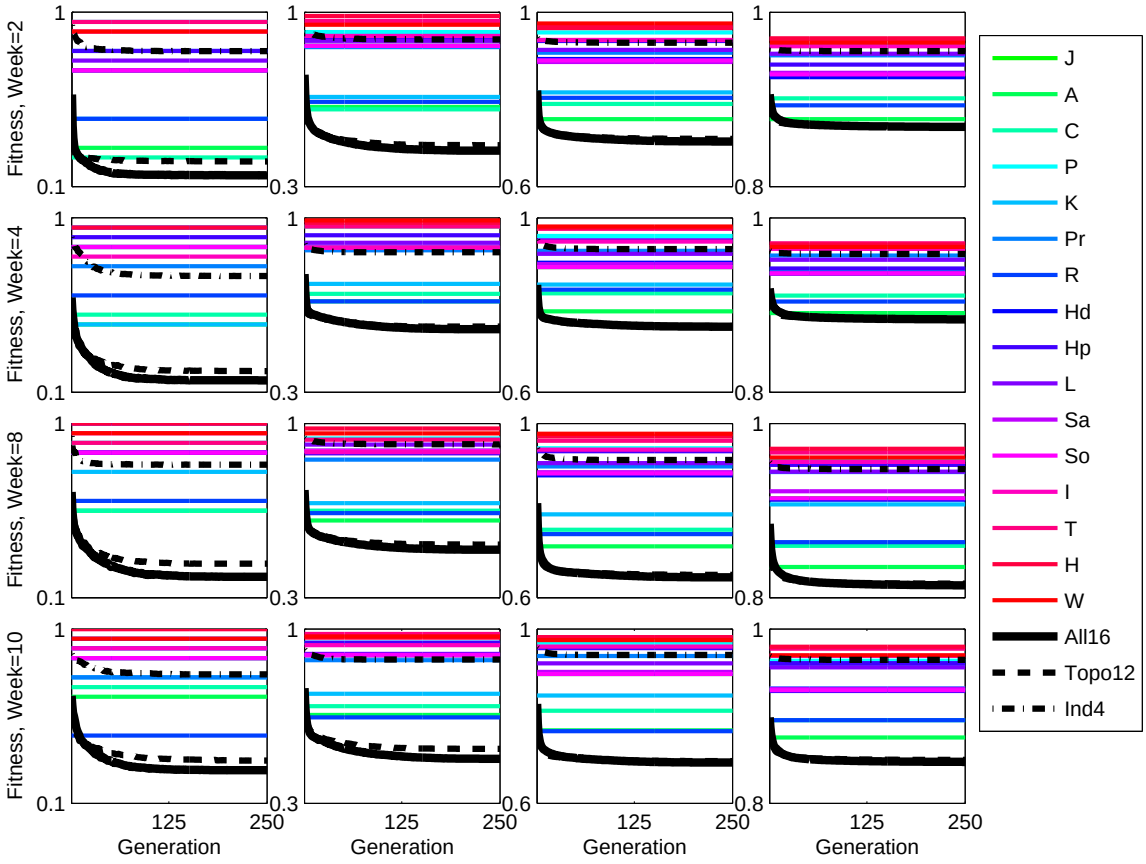
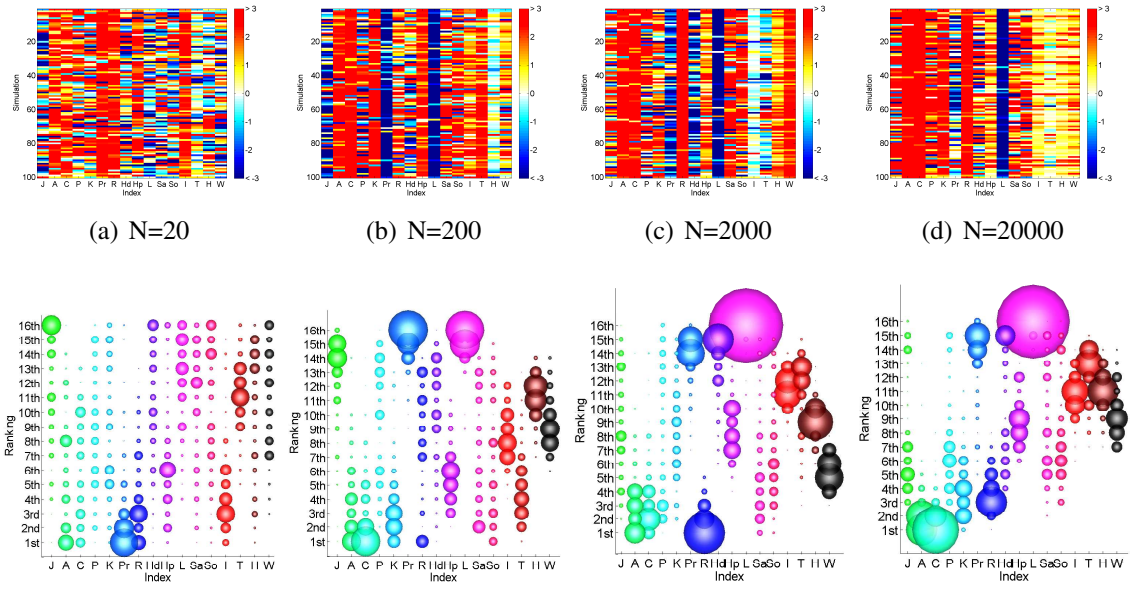


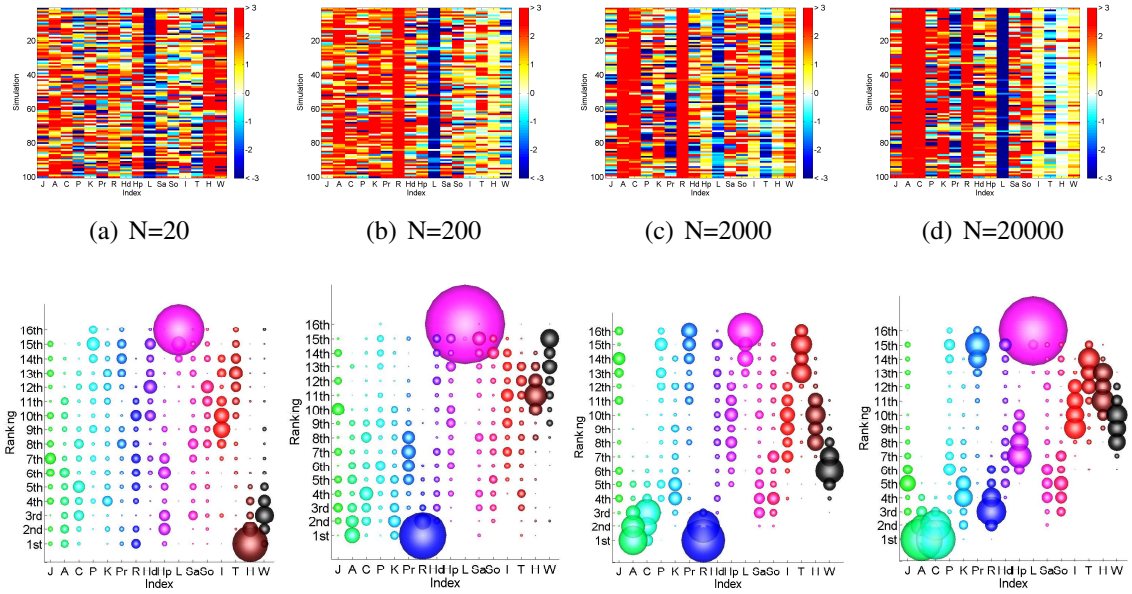
Figure 3.A1: Mean fitness computed from 100 simulations of CMA-ES. Training occurs on the new links that occur in a given week for each of (columns left to right) top N=20, top N=200, top N=2000 and top N=20,000. We compare each individual index, along with "all16" (evolved predictor consisting of all 16 indices), "topo12" (evolved predictor consisting of only the 12 topological indices), and "node4" (evolved predictor consisting of only the 4 node similarity indices). To show detail, the axes are not uniformly scaled between each column.



(a) N=20 (b) N=200 (c) N=2000 (d) N=20000

(e) N=20 (f) N=200 (g) N=2000 (h) N=20000

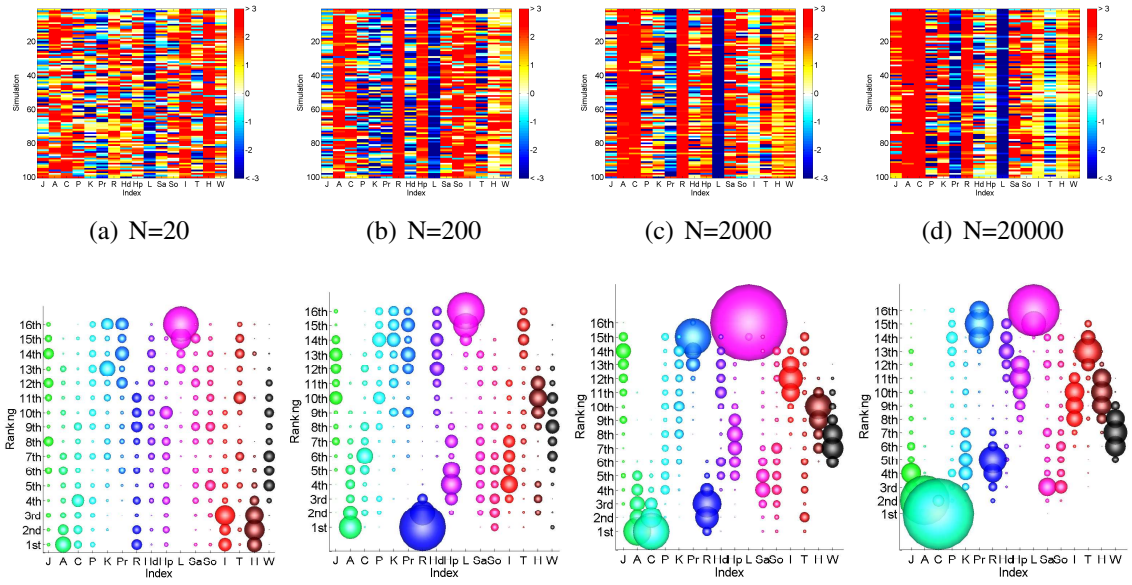
Figure 3.A2: Ranking of the value of the evolved coefficients from each of 100 CMA-ES runs, Weeks 1-2. Preferential attachment, resource allocation and common neighbors are the most frequently chosen top ranking (i.e., heavily weighted) indices. The lowest ranking index was LHN. Individual similarity indices, such as happiness, word similarity, Twitter user Id and Tweet count were ranked intermediate. J=Jaccard, A=Adamic-Adar, C=Common neighbors, P=Paths, K=Katz, Pr=Preferential attachment, R=Resource allocation, Hd=Hub depressed, Hp=Hub promoted, L=Leicht-Holme-Newman, Sa=Salton, So=Sorenson, I=Twitter Id similarity, T=Tweet count similarity, H=Happiness similarity, W=word similarity.



(a) N=20 (b) N=200 (c) N=2000 (d) N=20000

(e) N=20 (f) N=200 (g) N=2000 (h) N=20000

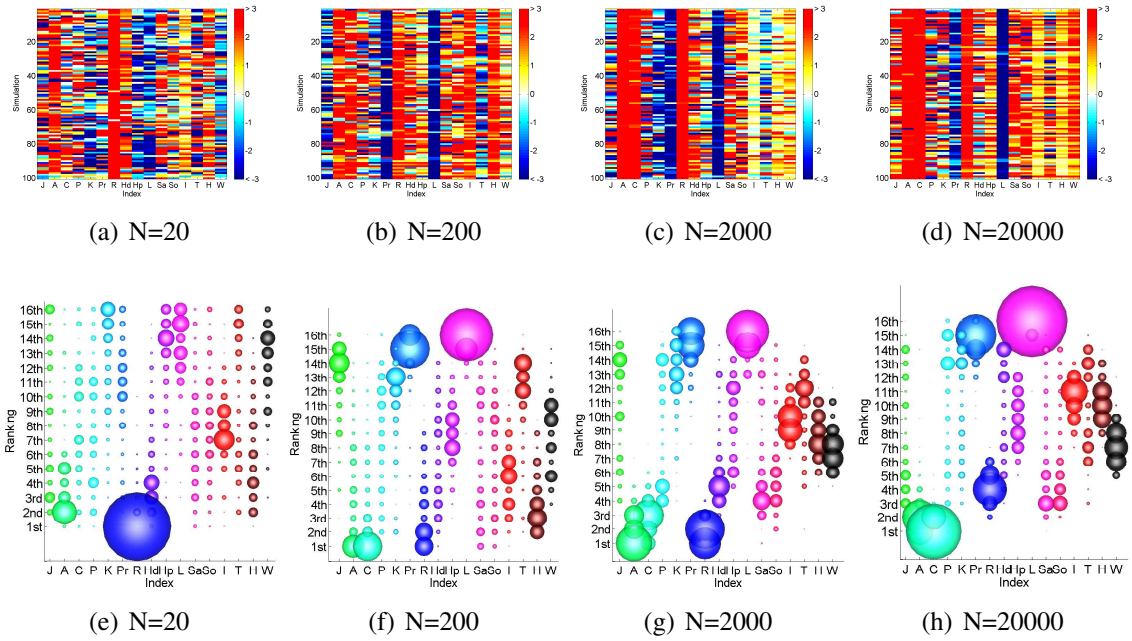
Figure 3.A3: Ranking of the value of the evolved coefficients from each of 100 CMA-ES runs, Weeks 3-4. Happiness similarity, resource allocation and Adamic-Adar are the most frequently chosen top ranking (i.e., heavily weighted) indices. The lowest ranking index was LHN. Individual similarity indices, such as happiness, word similarity, Twitter user Id and Tweet count were ranked intermediate. J=Jaccard, A=Adamic-Adar, C=Common neighbors, P=Paths, K=Katz, Pr=Preferential attachment, R=Resource allocation, Hd=Hub depressed, Hp=Hub promoted, L=Leicht-Holme-Newman, Sa=Salton, So=Sorenson, I=Twitter Id similarity, T=Tweet count similarity, H=Happiness similarity, W=word similarity.



(a) N=20 (b) N=200 (c) N=2000 (d) N=20000

(e) N=20 (f) N=200 (g) N=2000 (h) N=20000

Figure 3.A4: Ranking of the value of the evolved coefficients from each of 100 CMA-ES runs, Weeks 7-8. Happiness similarity, common neighbors and resource allocation are the most frequently chosen top ranking (i.e., heavily weighted) indices. The lowest ranking index was LHN. Individual similarity indices, such as happiness, word similarity, Twitter user Id and Tweet count were ranked intermediate. J=Jaccard, A=Adamic-Adar, C=Common neighbors, P=Paths, K=Katz, Pr=Preferential attachment, R=Resource allocation, Hd=Hub depressed, Hp=Hub promoted, L=Leicht-Holme-Newman, Sa=Salton, So=Sorenson, I=Twitter Id similarity, T=Tweet count similarity, H=Happiness similarity, W=word similarity.



(a) N=20 (b) N=200 (c) N=2000 (d) N=20000

(e) N=20 (f) N=200 (g) N=2000 (h) N=20000

Figure 3.A5: Ranking of the value of the evolved coefficients from each of 100 CMA-ES runs, Weeks 9-10. Resource allocation and common neighbors are the most frequently chosen top ranking (i.e., heavily weighted) indices. The lowest ranking index was LHN. Individual similarity indices, such as happiness, word similarity, Twitter user Id and Tweet count were ranked intermediate. J=Jaccard, A=Adamic-Adar, C=Common neighbors, P=Paths, K=Katz, Pr=Preferential attachment, R=Resource allocation, Hd=Hub depressed, Hp=Hub promoted, L=Leicht-Holme-Newman, Sa=Salton, So=Sorenson, I=Twitter Id similarity, T=Tweet count similarity, H=Happiness similarity, W=word similarity.

Chapter 4

Estimation of global network statistics from incomplete data

Complex networks underlie a variety of social, biological, physical, and virtual systems. In many settings, it is impossible to observe all nodes and all network interactions. Previous work addressing the impacts of partial network data, which is surprisingly limited and focuses primarily on missing nodes, suggests that network statistics derived from subsampled data are not suitable plug in estimators for network statistics describing the overall network topology. Our aim is to generate scaling methods to predict true network parameters from only partial knowledge of nodes, links, or weights. We validate analytical results on four simulated network classes (Erdös-Rényi, Scale-free, Small World, and Range dependent networks) each with $N = 2 \times 10^5$ and $k_{\text{avg}} = 10$ and empirical data sets of various sizes. We perform 100 subsampling experiments by varying proportions of sampled data and demonstrate that our scaling methods provide very good estimates of the true network parameters. Lastly, we apply our techniques to a set of rich and evolving large-scale social networks, Twitter reply networks. From over 100 million tweets, we use our scaling techniques to propose a statistical characterization of the Twitter interactome from September 2008-February 2009.

4.1 Introduction

Complex networks have been used to represent a variety of interactions in biological, social, and virtual realms. In practice, data collected about networks is often incomplete due to covert interactions or constraints in sampling. Particular individuals may wish to remain hidden, such as members of organized crime, and individuals who are otherwise overt may have some interactions that they wish to remain hidden because those interactions are of a sensitive nature (e.g., sexual contacts). In other instances, sampling constraints for extremely large networks necessitate an understanding of how network statistics scale under various sampling regimes [1, 2]. Explorations of empirically studied networks have largely ignored these biases and consequently, characterizations of the observable (sub)networks have been reported as if they characterize the "true" network of interest.

When members of a population are drawn at random, each with equal selection probability, the sample parameter being studied is often a good estimate of the population parameter. Problematically, global statistics of subnetwork data are often not good characterizations of the true network because subsamples can be biased in that some individuals or interactions may be more likely to be selected in a subsample [3]. As an example, consider a network for which a random selection of links are observed. The collection of observed nodes in such a subnetwork is biased because large degree nodes are more likely to be included in the sample than nodes of small degree.

The errors introduced by biases in sampling may be exacerbated by particular sampling strategies and also by various underlying network topologies of the true network from which the subsamples are chosen [4–11]. Kossinets [7] highlighted the missing data problem for social networks and demonstrated that both the sampling strategy and underlying

network topology influence how a particular network parameter scales. Kossinets documents that subnetwork statistics are often not good estimates of true network statistics, in some cases, producing an error of over 200% [7]. Other researchers have similarly explored sampling by nodes [1, 5, 9, 12–14], sampled edges or messages [1, 2, 14], and graph exploration methods based on random walks, snowball sampling or respondent driven sampling [1, 15, 16]. Others have explored biases in these sampling regimes [14–19].

The development of techniques to correct sample estimates of population parameters would enable more accurate portrayals of empirically studied networks and aid in efforts to model cascading failures, as well as complex contagion. Additionally, in cases where a characterization of a large network is desired, the development of techniques to probe a system with minimal effort and computational resources would greatly aid in the understanding of large network data sets. Before proceeding, we outline some of the most common global network statistics.

4.1.1 Global network statistics

Networks may be characterized by a variety of network statistics. In this paper, we explore how descriptive measures such as the number of nodes (N), the number of edges (M), the average degree (k_{avg}), clustering coefficient (C) [20], the proportion of nodes in the giant component (S) and the max degree ($k_{\max}$) scale with respect to missing network data and suggest predictor methods for inferring true network statistics from subsampled network data. Problematically, sample statistics are not good estimators for the true, underlying network from which the sample was drawn. Relatively few studies focusing on how missing network data impacts inferred network topology provide analytical results that can be applied to scale subsampled network statistics to values described the full, often unknown,

network. Those that have provided analytical results focus primarily on subnetworks induced by sampled nodes [12, 21].

In addition to these parameters, the degree distribution, $Pr(k)$, can provide valuable insight into network structure. The classical Erdös-Rényi random graph growth model exhibits a Poisson degree distribution, $Pr(k) = \frac{\lambda^k e^{-\lambda}}{k!}$ [22]. In contrast to Erdös-Rényi random networks, preferential attachment growth models describe a random process whereby new nodes attach preferentially to nodes of large degree giving rise to a Powerlaw or Scale-free degree distribution, $Pr(k) \propto k^{-\gamma}$ [23–26]. Other distributions, such as lognormals and powerlaws with exponential cutoffs may equally characterize the degree distributions of some empirical networks [27].

Previous work has explored how the degree distribution is distorted when the subnetwork is the induced subgraph on sampled nodes [5, 6, 9, 13, 14, 28–30]. Han et al. [5] investigate the effect of sampling on four types of simulated networks (random graphs with (1) Poisson, (2) Exponential (3) Power-law and (4) Truncated normal distributions). They observe that degree distributions of sampled Erdös-Rényi random graphs appear to be linear on a log-log plot. Others have also suggested that subnetworks of Erdös-Rényi random graphs appear "powerlaw-like" and could be mistaken for a scale-free network [5, 13]. Typically, scale-free networks have degree distributions which span several orders of magnitude and thus, subnetworks of Erdös-Rényi random graphs would not be classified as scale-free networks by most researchers. As warned by Clauset, Shalizi and Newman [27], further errors may be incurred when attempting to use linear regression to fit a power-law.

Stumpf and Wiuf [28] examine how degree distributions of Erdös-Rényi random graphs scale when subnetworks are obtained through uniform random sampling on nodes and "preferential sampling of nodes," whereby large degree nodes have a greater probability

of being selected. They show that Erdös-Rényi random graphs are closed under subsampling by nodes, but not under preferential sampling of nodes.

Stumpf et al. [9] suggest that the degree distribution of the subnetwork induced on randomly selecting nodes is independent of the proportion of nodes sampled and that the true degree distribution can only be determined by knowledge of the generating mechanism for the network. Problematically, this is often not known or fully understood.

To understand these distortions, we consider the probability that a node v will have degree k in a subnetwork. Under uniform random sampling of nodes, a node of degree i in the true network will become a node of degree k in the subnetwork ($k \leq i$) with probability $Pr(k|i) = \binom{i}{k} q^k (1-q)^{i-k}$. The subnetwork degree distribution can be determined by weighting these probabilities by $Pr(i)$, the probability of node i appearing in the true network [31]. The subnetwork degree distribution is given by

$$\tilde{Pr}(k) = \sum_{i=k}^{k_{\max}} \binom{i}{k} q^k (1-q)^{i-k} Pr(i). \tag{4.1}$$

Several researchers have explored techniques for estimating the true degree distribution from subnetwork data. One approach involves viewing Equation 4.1 as a system of equations and solving for the true degree distribution in terms of the (observed) subnetwork degree distribution. Denoting $\hat{Pr}(k)$ as the predicted degree distribution yields

$$\hat{P}_r(k) = \sum_{i=k}^{k_{\max}} (-1)^{i-k} \binom{i}{k} q^{-i} (1-q)^{i-k} \tilde{P}_r(i). \tag{4.2}$$

Our result is similar[1] to the derivation provided by Frank [29],

$$\hat{P}_r(k) = \sum_{i=k} (-1)^{i-k} \binom{i}{k} q^{-i-1} (1-q)^{i-k} \tilde{P}_r(i), \tag{4.3}$$

which is not guaranteed to be non-negative [3].

Model selection methods employ maximum likelihood estimates to select which type of degree distribution characterizes a true network, given only a subnetwork degree distribution [32]. Although this method highlights that some heavy tailed empirical networks may be better characterized by lognormal or exponential cutoff models instead of power laws, the shortcoming of this method is that only models selected *a priori* for testing form the candidate pool of possible distributions.

In contrast to the model selection technique proposed by Stumpf et al. [32], we explore a probabilistic approach which utilizes knowledge of the proportion of sampled network data (q) and the subnetwork degree distribution. In doing so, we desire an estimation that captures the qualitative nature of the degree distribution without making any assumptions about candidate models. We show that reasonably good estimations of $Pr(k)$ can be achieved with no knowledge of the generating mechanism. With a reasonable estimation for the degree distribution available, we are able to overcome a previously noted obstacle identified by Kolaczyk [3]. He notes that predictors for network statistics when sampling by links has proven more elusive because of the need for knowledge of the true degree distribution [3]. Our estimations can be used in conjunction with Hortiz-Thompson estimators to reasonably predict network statistics for cases where node selection is not uniform (i.e., subnetworks obtained by the induced subgraph on sampled links or weights).

[1] Our derivation differs from [29] by a factor of q. When $k = k_{\max}$, our result becomes $\hat{Pr}_{k_{\max}} = \sum_{i=k_{\max}}^{k_{\max}^{\text{obs}}} (-1)^{i-k} \binom{i}{k} q^{-i} (1-q)^{i-k} P_i = (-1)^0 \binom{k_{\max}}{k_{\max}} q^{-k_{\max}} (1-q)^0 P_{k_{\max}}$. This supports our derivation (Equation (4.2)).

In the subsequent sections, we summarize this work and show how our method surmounts this obstacle. To our knowledge, scaling techniques for networks generated by sampled by interactions (e.g., weighted networks) have not been addressed in the literature and given the interest in large, social networks derived from weighted, directed interactions, we find this analysis timely and relevant.

4.2 Sampling techniques and missing data

In this paper, we focus on four sampling regimes (1) subnetworks induced on randomly selected nodes, (2) subnetworks obtained by random failure of links, (3) subneworks generated by randomly selected links, and (4) weighted subnetworks generated by randomly selecting interactions. Motivated by our work with Twitter reply networks [33] for which we have a very good approximation of the percent of messages which are obtained, we base our work on the assumption that the proportion of missing data is known. This is a critical assumption and one that we acknowledge may not always be satisfied in practice. Efforts to estimate the proportion of missing nodes or links are intriguing, but are beyond the scope of this paper.

The remainder of this paper is organized as follows. In Section 4.3, we describe our data. In Section 4.4, we describe our sampling strategies in greater detail and describe scaling methods for global network statistics. We apply our methods to four classes of simulated networks and six empirical datasets. In Section 4.5 we apply our methods to Twitter reply networks as both a case of scientific interest and demonstration of our methods. In Section 4.6, we discuss the implications of these findings and suggest further areas of research.

4.3 Methods

4.3.1 Unweighted, undirected networks

Our data consist of simulated and empirical networks. We generate unweighted, undirected networks with $N = 2 \times 10^5$ and $k_{\text{avg}} = 10$ according to four known topologies: Erdös-Rényi random graphs with a Poisson degree distribution [22], Scale-Free random graphs with a power law degree distribution [24, 34], Small world networks [35], and Range dependent networks [36] .[2] We also examine six well known empirical network datasets: *C. elegans* [35, 38], Airlines [39], Karate Club [40], Dolphins [41], Condensed matter [42], and Powergrid [35]).

Each of these simulated and empirical networks is subsampled and the subnetwork is taken to be the subnetwork induced on sampled nodes (Fig. 4.1), the subnetwork obtained by failing links (Fig. 4.2), or the subnetwork generated by sampled links (Fig. 4.3). For a given network, 100 simulated subnetworks are obtained for a given sampling strategy and given q as q varies from 5% to 100% (in increments of 5%).

[2]Erdös-Rényi, Scale-free, Small world and Range dependent models were constructed with the CONTEST Toolbox for Matlab [37]. We note that the small world networks were set to have random rewiring probability $p = 0.1$ and preferential attachment networks were set to have $d = 5$ new links when they enter the network. Range dependent networks were set to establish a link between nodes v_i and v_j with probability $\alpha\lambda^{|j-i|-1}$ where we set $\lambda = 0.9$ and $\alpha = 1$. As noted by [37], this choice of α ensures that nodes v_i and v_{i+1} are adjacent and $\lambda^{|j-i|-1}$ ensures that short range connections are more probable than long range connections.

4.3.2 Weighted, undirected networks

We examine the effects of uniformly increasing edge weight (Experiment 1, Cases I-V) as well as the distribution of edge weights (Experiment 2, Cases VI and VII) on the scaling of network statistics (Table 4.1).

Table 4.1: Summary of weighted network experiments. Note: $w(e_j)$ refers to the weight of edge e_j, $s(v_j)$ refers to the strength of node v_i) and $randi\,\{1..9\}$ refers to a randomly selected integers between 1 and 9 (inclusive).

Case	k_{avg}	w_{avg}	Distribution of weights
I	6	1.0	$w(e_j) = w_{\text{avg}}$ (uniform)
II	6	2.0	$w(e_j) = w_{\text{avg}}$ (uniform)
III	6	3.0	$w(e_j) = w_{\text{avg}}$ (uniform)
IV	6	4.0	$w(e_j) = w_{\text{avg}}$ (uniform)
V	6	5.0	$w(e_j) = w_{\text{avg}}$ (uniform)
VI	6	5.0	$s(v_i) = \left\lceil \frac{30}{k} \right\rceil$ (equal effort)
VII	6	5.0	$w(e_j) = randi\,\{1..9\}$ (randomized)

Experiment 1: Uniform distribution of edge weights

In this set of experiments we generate Erdös-Rényi networks with $N = 2000$ nodes and $k_{\text{avg}} = 6$. Each edge receives equal weight, w, where $w = 1, 2, 3, 4,$ or 5 (corresponding to Cases I-V). We similarly generate Scale-free networks with $N = 2000$ nodes and $k_{\text{avg}} = 6$. Each of the weighted, undirected networks described is subsampled by randomly selecting $q \sum_{e_i \in E(G)} w(e_i)$ interactions. The subnetwork is taken to be the network generated by links with $w(e_j) > 0$ (Fig. 4.4). One hundred subnetworks are obtained for each network for varying proportions of sampled interactions (q).

Experiment 2: Non-uniform distribution of edge weights

In this set of experiments, we explore how the distribution of weights on edges can impact scaling of global network statistics. As in the previous case, we first generate an Erdös-Rényi network with $N = 2000$ and $k_{\text{avg}} = 6$. We then add weights to edges in one of two ways. In Case VI, we assume "equal effort" in that all nodes will have an equal number of interactions distributed equally among their incident edges. This requirement ensures that all nodes have equal node strength and that effort is equally distributed to each neighbor. More specifically, for node $\deg(v_i) = k$, we set each of the k edges to have weight $\lceil \frac{30}{k} \rceil$. In Case VII, for each edge we select an integer weight between 1 and 9 from a uniform probability distribution. Certainly, other variants of the weight distribution exist and their analysis may provide additional insight in future studies.

4.3.3 Weighted, directed networks - Twitter reply networks

Twitter reply networks [33] are weighted, directed networks constructed by establishing a directed edge between two users if we have a directed reply from a user to another during the week under analysis. These networks are derived from over 100 million tweets obtained from the Twitter API service during September 2008 to February 2009.[3] During this time, we obtained between 25% to 55% of all tweets 4.A24. Using the scaling methods developed in Sections 4.4.1-4.4.4, we predict global network statistics for the Twitter interactome during this period of time by viewing in- and out-network statistics separately (e.g., two distinct networks) to account for directionality.

[3] We refer the interested reader to [33] for more information.

4.4 Estimating global network statistics

4.4.1 Sampling by nodes

Given a network, $G = (V, E)$, where V is the collection of nodes (or vertices) and E is the collection of links (or edges), we randomly select a portion of nodes q, where $0 < q \leq 1$. The node induced subgraph on these randomly sampled nodes is given by $G^* = (V^*, E^*)$, where V^* represents the randomly selected nodes and E^* represents the edges in E for whom both endpoints lie in V^* (Fig. 4.1). This type of sampling occurs when a selected group, representative of the whole, is observed and all interactions between sampled individuals are known. This sampling strategy is well studied and we will overview key results here (see [3]).

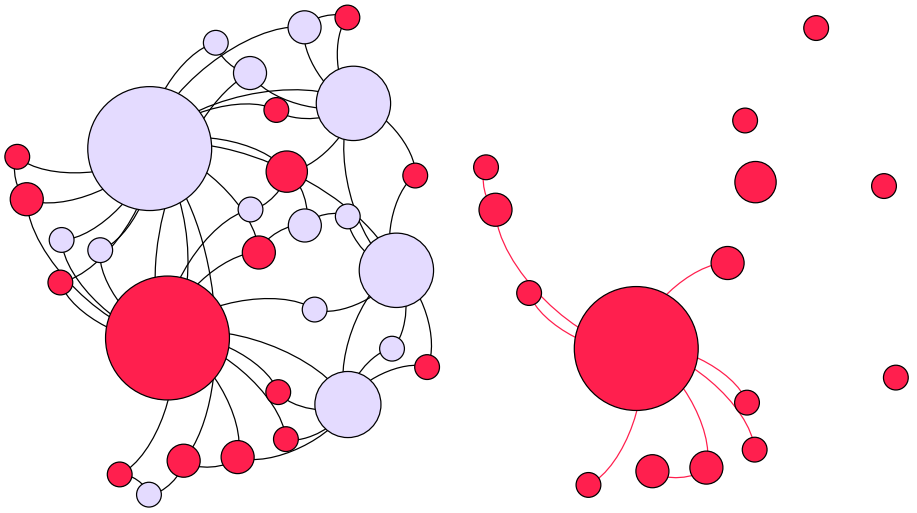
Scaling of $N, M, k_{\text{avg}}, C, k_{\max}, S$

Given a subnetwork of size $n = qN$ known to be obtained by randomly selecting qN nodes, the number of nodes in the subsample clearly scales linearly with q (Figs. 4.A1a and 4.A2a). The size of the true network is predicted by

$$\hat{N} = \frac{1}{q} n, \tag{4.4}$$

which shows good agreement with the true network parameter (Table 4.A1). Note that this result is independent of network type and is only dependent on q, the fraction of nodes subsampled, and n, the size of the subsample.

Given a network with N nodes and a subnetwork of n nodes, the probability of selecting edge e_{ij} is given by $\frac{n(n-1)}{N(N-1)}$. This is simply the probability that the two nodes, v_i and v_j,



(a) Sampled nodes (b) Nodes induced subnetwork

Figure 4.1: Node induced subnetwork on randomly sampled nodes. (a) The true network is sampled by randomly selecting nodes (red). (b) The node induced subnetwork consists of sampled nodes and edges whose endpoints both lie in the collection of sampled nodes.

incident with the edge e_{ij}, are selected. The number of edges in the subnetwork is found by

$$m = \frac{n(n-1)}{N(N-1)} \cdot M, \tag{4.5}$$

where m represents the number of edges in the subnetwork and M represents the number of edges in the true network. For large networks, $m \approx q^2 M$. This agrees well with simulated results (Figs. 4.A1b and 4.A2b). The predicted number of edges is given by

$$\hat{M} = m \cdot \frac{N(N-1)}{n(n-1)}, \tag{4.6}$$

which scales as $\hat{M} \approx \frac{1}{q^2}m$ for large networks. This predictor shows good agreement with actual values (Table 4.A2).

The average degree, k_{avg}, is found by

$$k_{\text{avg}} = \frac{2M}{N}.$$

Given expressions for the expected number of edges (4.6) and the expected number of nodes (4.4), the expected average degree of a true network, $\hat{k}_{\text{avg}}$, based on an observed average degree of a subnetwork:

$$\begin{aligned}
\hat{k}_{\text{avg}} &= \frac{2\hat{M}}{\hat{N}} && (4.7)\\
&= \frac{2m \cdot \frac{N(N-1)}{n(n-1)}}{\frac{n}{q}} && (4.8)\\
&= \frac{2m}{n} \cdot \frac{N-1}{n-1} && (4.9)\\
&= k_{\text{avg}}^{\text{obs}} \cdot \frac{N-1}{n-1} && (4.10)\\
&\approx \frac{k_{\text{avg}}^{\text{obs}}}{q}, && (4.11)
\end{aligned}$$

where in line (10) we have assumed that $\hat{N} \approx N$, $N >> 1$ and $n >> 1$. Comparing this result to simulated subnetworks induced by subsampling nodes (Figs. 4.A1c and 4.A2c), we find very good agreement between the predicted average degree and true average degree (Table 4.A3), except for the small empirical networks (Karate club and Dolphins) sampled with low q. In these cases, we violate the assumption that $n >> 1$ because subsamples of the Karate Club network degenerate to subnetworks of 3 edges or less when $q \leq 0.20$.

Similarly, subsamples of the Dolphin network degenerate to subnetworks of 3 edges of less when $q \leq 0.15$. When the observed number of edges in the subnetwork exceeds 3, our predicted $\hat{M}$ has an error less than 5% (Table 4.A3).

The scaling of the max degree is highly dependent on network type, or more precisely, the relative frequency of high degree nodes. For networks with relatively few large hubs and many small nodes of small degree, $k_{\max}$ scales linearly with q and $\hat{k}_{\max} \approx \frac{k_{\max}}{q}$. For networks with many nodes of maximal degree[4] $k_{\max}$ scales nonlinearly with q (Figs. 4.A1d and 4.A2d).

This distinction makes predicting the maximum degree more challenging since an accurate predictor ultimately relies on knowledge of the network type - knowledge one usually does not have in an empirical setting. Our proposed technique utilizes $\hat{k}_{\max} \approx \frac{k_{\max}^{obs}}{q}$, unless our algorithm detects a large number of nodes with degree similar to $k_{\max}$ and are assured that the subnetwork that has not degenerated to a small network ($n < 30$).[5] In this case,

$$\hat{k}_{\max} \approx \frac{k_{\max}^{obs}}{1 - \frac{q}{\theta}}, \tag{4.12}$$

where θ =the number of nodes with degree greater than 75% of $k_{\max}$. The rationale for this rough approximation is that the nodes which have high degree ($> 75\%$ of the observed max. degree) may have been nearly equal contenders for losing a neighbor during subsampling. When all nodes have equal degree, the denominator of Equation 4.12 tends to $\hat{k}_{\max} \approx k_{\max}^{obs}$. Table 4.A4) presents the error for this predictor and demonstrates that our method performs

[4] An example of this would be a regular lattice. All nodes have the same (and hence maximal) degree. This pathological example is not often seen in practice. Simulated Small world networks begin as a regular lattice with random rewiring probability, p. Since our Small world networks have $p = 0.1$, our Small world networks exhibit this pathological behavior more so than several empirical Small world networks. We note that this is simply a matter of tuning p and not indicative of all Small world networks.

[5] More specifically, if our algorithm detects $n_{k_{\max}-1} \cdot k_{\max} - 1 > n_{k_{\max}} \cdot k_{\max}$, then we use the adjustment Equation 4.12, where $n_{k_{\max}-1}$ represents the number of nodes of degree $k_{\max} - 1$.

reasonably well for most network in our data set. To our knowledge, this is the first attempt to characterize how $k_{\max}$ scales with subsampling and we hope that future work improves upon our estimate.

We measure clustering using Newman's global clustering coefficient [20] $C_G = \frac{3\times\tau_\Delta(G)}{\tau_3^+(G)}$, where $\tau_\Delta(G)$ denote the number of triangles on a graph and $\tau_3^+(G) = \tau_3(G) - 3\tau_\Delta(G)$, which is the number of vertex triples connected by exactly two edges (as in the notation used by [3]). Since the probability of selecting a node is q, both the number of triangles and connected vertex triples scale as q^3. Thus, $\hat{\tau}_\Delta(G) = \frac{1}{q^3}\tau_\Delta(G^*)$ and $\hat{\tau}_3^+(G) = \frac{1}{q^3}\tau_3^+(G^*)$ [21]. We then expect

$$\hat{C}_G \approx C_G^*. \tag{4.13}$$

This is supported by simulations (Figs. 4.A1e and 4.A2e) and small errors in $\hat{C}_G$ (Table 4.A5). We note that for small q, some subnetworks completely breakdown and no connected triples are present. In these situations, the clustering coefficient can not be computed nor can the true network's clustering coefficient be well predicted.

We next explore how the size of the giant component scales with the proportion of nodes sampled (Fig. 4.A1f and 4.A2f). For the Erdös-Rényi and Scale-free random graphs, the giant component emerges when the subnetwork has $k_{\text{avg}}^{sub} > 1$. This occurs when $qk_{\text{avg}} > 1$ and so for our simulated Erdös-Rényi and Scale-free networks, this occurs when $q = .10$ because the true networks have $k_{\text{avg}} \approx 10$. The thresholds for the emergence of the giant component in Small World and Range dependent networks are much higher. This may be due to the relatively large clustering coefficients of these networks. As suggested by Holme et al. [43], networks with a large (Watts and Strogatz [35]) clustering coefficient are more

vulnerable to random removal of nodes. We observe the same trend with Newman's global clustering coefficient.

In the case of the empirical networks, we find that the giant component emerges for q corresponding to $k_{\text{avg}}^{obs} > 1$. *C. elegans*, Airlines and Condensed Matter networks are more resilient to random removal of nodes in that the giant component persists for small levels of q. This is most likely due to their relatively high average degrees, as compared to the other networks (heterogeneity of nodes' degrees in these networks). Heterogeneous networks demonstrate more resilience due to random removal of nodes at high levels of damage [44]. In general, it may be very difficult to predict the exact critical point at which the giant component emerges from subnetwork datasets.

Scaling of $Pr(k)$

The complementary cumulative degree distribution (CCDF) becomes more distorted as smaller proportions of nodes are sampled, as shown in Figure 4.A3 and given by Equation 4.1. Subnetworks obtained by the induced graph on sampled nodes will often have $\tilde{Pr}(0) > 0$. This occurs when v_i is selected in sampling, but no neighbors of v_i are selected in the sample.

Our goal is to predict the degree distribution, given only knowledge of the proportion of nodes sampled (q) and the subnet degree distribution. We note that the probability that an observed node of degree k came from a node of degree $j \geq k$ in the true network is given by

$$Pr(k|j) = \begin{cases} \binom{j}{k} q^k (1-q)^{j-k}, & \text{when } j \geq k \\ 0, & \text{when } j < k, \end{cases}$$

where q is the probability that a node's neighbor was included in the subsample and $1-q$ is the probability that a node's neighbor is not included in the subsample.

After normalizing, we find $\psi(j) = \frac{Pr(k|j)}{c}$ describes the normalized probability that an observed node of degree k came from a node of degree j in the true network, where $c = \sum_{j=k}^{\infty} Pr(k|j)$. Note that when $|1-q| < 1$ this series converges and we find $c = \sum_{j=k}^{\infty} Pr(k|j) = \frac{1}{q}$. Thus,

$$\psi(j) = \begin{cases} q\binom{j}{k}q^k(1-q)^{j-k}, & \text{when } j \geq k \\ 0, & \text{when } j < k, \end{cases} \tag{4.14}$$

Considering all nodes of degree, n_k, we compute

$$n_k \cdot \psi(k) = n_k \cdot \left(\frac{\binom{j}{k}q^k(1-q)^{j-k}}{c} \right) \tag{4.15}$$

$$= n_k \cdot \left(q\binom{j}{k}q^k(1-q)^{j-k} \right), \tag{4.16}$$

where Stirling's approximation is used to approximate the binomial coefficients for large j. Care is taken to include observed nodes of degree zero in this process (e.g., $k = 0$ in Equation 4.15).

For networks with large degrees (e.g., hubs), one can further speed up the computation and reduce floating point arithmetic errors by mapping back observed nodes of degree k to

the expected value of the distribution obtained in Equation 4.14:

$$E(j) = \frac{1}{c}\sum_{j=k}^{\infty} j\binom{j}{k}q^k(1-q)^{j-k} \tag{4.17}$$

$$= q\frac{1-q+k}{q^2} \tag{4.18}$$

$$\approx \frac{k}{q}, \text{ for } k >> 1, \tag{4.19}$$

where $c \approx \frac{1}{q}$. In making use of Equation 4.17, we perform a separate calculation for nodes of degree zero: $\left\{n_0 \cdot \frac{(1-q)^j}{\sum_j (1-q)^j}\right\}_{j=1}^{4k_{\max}^{obs}}$.[6]

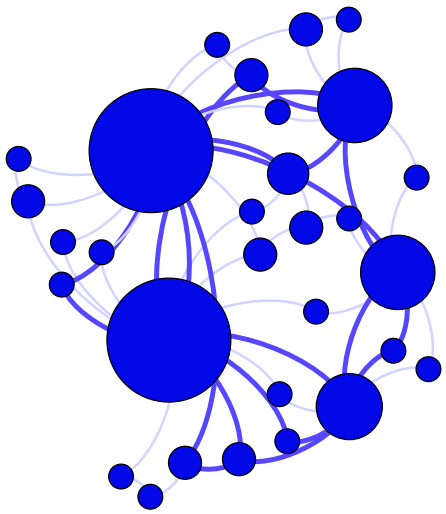
Figure 4.A4 reveals the predicted degree distribution for subnets induced on varying levels of randomly selected nodes. To test the goodness of fit for the estimated degree distribution and the true $Pr(k)$, we apply the two sample Kolmogorov-Smirnov test. Figure 4.A16 shows the D test statistics for the predicted degree distributions for both estimation methods (Equations 4.15 and 4.17), as well as the D_{crit} computed from $c(\alpha)\sqrt{\frac{n_1+n_2}{n_1 n_2}}$, where $c(0.05) = 1.36, n_1 = k_{\max}$ and $n_2 = \hat{k}_{\max}$. For most networks, $D \leq D_{\text{crit}}$ for $q \geq 0.3$, suggesting that when at least 30% of network nodes are sampled, our methods provide an estimated degree distribution which is statistically indistinguishable from the true degree distribution. Although we reject the null hypothesis for the preferential attachment case, for all $q \neq 1$, we wish to point out the potential for bias in the Kolmogorov-Smirnov test with large n [45]. As shown, D_{crit} values are quite low and the bias in this test is due to large n_1, n_2. The statistical power in this test leads to the detection statistically significant differences, even when the absolute difference is negligible. Thus, we caution the interpretation of this statistical test and place more interest in the value

[6] In all cases, we assume a finite network. We limit our calculations to $4 \cdot k_{\max}^{obs}$ as a rough estimate on the upper bound needed for the sum in Equation 4.14.

$D = \max |F_{i,true} - F_{i,predicted}|$, where F_{true} and $F_{prediction}$ represent the true and predicted CDFs.

4.4.2 Link failure

We know turn our attention to link failure. As in the previous cases, we denote the true, unsampled network as $G = (V, E)$. Some proportion, q of links remain "on" (or present in the sample) and $1 - q$ are hidden or undetected by sampling. $E^* \subseteq E$ consists of precisely the links that remain "on" and $V^* = V$ (Fig. 4.2).



(a) Failed links

Figure 4.2: Failed link subnetwork. Hidden or missing links are depicted in grey. All nodes remain in the subnetwork and only visible or sampled links remain.

In this case we may use the plug-in estimator to predict the number of nodes, $\hat{N} = n$ and we may predict the number of edges by $\hat{M} = \frac{m}{q}$. The average degree is found by

$$\hat{k}_{\text{avg}} = \frac{2\hat{M}}{\hat{N}} \tag{4.20}$$

$$= \frac{2m}{qn} \tag{4.21}$$

$$= \frac{k_{\text{avg}}^{\text{obs}}}{q}. \tag{4.22}$$

Newman's global clustering coefficient $C_G = \frac{3\times\tau_\Delta(G)}{\tau_3^+(G)}$ [20] and note that $q^3\tau_\Delta(G) = \tau_\Delta(G^*)$ and $q^2\tau_3^+(G) = \tau_3^+(G^*)$ because each edge is selected with probability q. Thus,

$$\begin{aligned} C_G^* &= \frac{3 \times \tau_\Delta(G^*)}{\tau_3^+(G^*)} \\ &= \frac{3q^3 \times \tau_\Delta(G)}{q^2\tau_3^+(G)} \\ &= qC_G. \end{aligned}$$

Thus,

$$\hat{C}_G = \frac{1}{q}C_G^*. \tag{4.23}$$

The maximum degree is computed with the same method as described in Section 4.4.1 because the number of neighbors of a node scales the same whether nodes or edges are removed from the network. Using these estimates, we find relatively low error in the predicted the network measures for $N, M, k_{\text{avg}}, k_{\max}$, and C_G (Tables 4.A6– 4.A10).

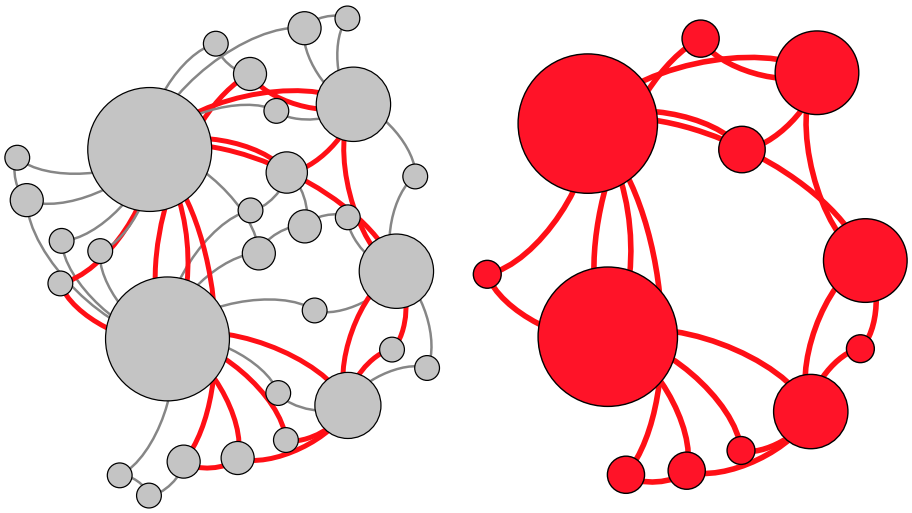
Several networks' giant component exhibit similar patterns of resilience when sampling by nodes or failing links. Comparing the resilience of the proportion of nodes in the giant

component under sampling by nodes vs. failing links, we see that Erdös-Rényi random graphs, random graphs with preferential attachment, Airlines, Condensed matter, *C. elegans* and Powergrid networks all perform relatively similar under the sampling regimes. A noticeable difference is seen in Small world, Range dependent, Karate club and dolphins. In the case of Small world and Range dependent, the regularity of the underlying lattice in these networks means that each time a node is not observed, this also means that k_{avg} edges are also missing. Given that the majority of nodes have the same degree for these networks these networks fracture the giant component quickly (i.e., for q around 0.7 and 0.8 respectively). In the case of the small Karate club and Dolphins networks sampled by nodes, the proportion of nodes in the giant component increases with decreasing q. In these cases, the network consists of relatively few nodes, which are connected. In contrast, when examining the failing links case, we have all nodes present but these nodes are missing almost all links and the network is highly disconnected.

Figure 4.A7 reveals the distortion of the CCDF when links fail in a network and all nodes remain known to the observer. Clearly, there are nodes of degree zero that are observed in this sampling regime. The predicted degree distribution is obtained by the methods described under sampling by nodes (including the treatment of observed nodes of degree zero). The results of the two sample test Kolmogorov-Smirnov reveal that the estimated degree distribution and the true degree distribution are statistically indistinguishable for $q \geq 0.3$ for most networks (Fig. 4.A17). As previously noted, the large number of observations in degree distribution for the random graph grown with preferential attachment leads to high statistical power and a low D_{crit}.

4.4.3 Sampling by links

The problem of missing links may also manifest itself in another manner. In contrast to the case when all nodes are known and some links are hidden links, we now consider subnetworks generated by sampled links and the nodes incident to those links (Fig. 4.3). This type of sampling occurs in many social network settings, such as networks constructed from sampled email exchanges or message board posts. In this case, we have data pertaining to messages (links) and nodes (individuals) are only discovered when a link (email) which connects to them is detected.



(a) Sampled links (b) Link induced subnetwork

Figure 4.3: Link induced subnetwork. (a) A network is sampled by randomly selecting links shown in red. (b) The subnetwork consists of all sampled links and only nodes which are incident with the sampled links. In this type of sampling, no nodes of degree zero are included in the network. Large degree nodes are more likely to be included in the subnetwork.

In this case, edges are sampled uniformly at random and we may use our previous estimator, $\hat{M} = \frac{m}{q}$. Node inclusion is biased, however, in that nodes of high degree will detected with greater probability than nodes of low degree precisely because they are more likely to have an incident edge sampled.

To motivate an appropriate predictor, we must first consider how the number of nodes in a subnetwork obtained by the subnetwork generated by sampled links scales with q (Figs. 4.A9a and 4.A10a). To do this, let us consider the probability that a node is included in such a subsample. If the number of edges not sampled ($M - m$) is less than the degree $k(v_i)$ of node v_i, then we can be certain that our node of interest will be detected in sampling. On the other hand, if $M - m \geq k(v_i)$, then the probability of v_i being in the subnetwork scales nonlinearly with q. Using the framework set forth by Kolaczyk [3], observe that there are $\binom{M-k}{m}$ ways of choosing m edges from the $M - k$ edges not incident with node v_i and there are $\binom{M}{m}$ total ways of choosing m edges from all M. Thus, we have

$$\begin{aligned} P(v_i \text{ is sampled}) &= 1 - P(\text{no edge incident to } v_i \text{ is sampled}) \\ &= \begin{cases} 1 - \frac{\binom{M-k(v_i)}{m}}{\binom{M}{m}}, & \text{if } m \leq M - k(v_i) \\ 1, & \text{if } m > M - k(v_i). \end{cases} \end{aligned}$$

The Horvitz-Thompson estimator given by

$$\hat{N} = \sum_{v_i \in V^*} \frac{1}{\pi_i}, \tag{4.24}$$

where $\pi_i = P(v_i \text{ is sampled})$.

Kolaczyk [3] warns that this may not be a useful result, due to the fact that the true degree of a given node is likely to be unknown. In our paper, we overcome this limitation by

using our predicted degree distributions obtained by the techniques previously mentioned. Observe that when sampling by links, no nodes of degree zero will be observed. We also note that in the case when $k << M$ and m, we may make the following approximation which is less computationally burdensome:

$$\begin{aligned}
\frac{\binom{M-k}{m}}{\binom{M}{m}} &= \frac{(M-k)!M-m)!}{M!(M-m-k)!} \\
&= \frac{(M-m)(M-m-1)(M-m-2)\dots(M-m-(k-1))}{M(M-1)(M-2)\dots(M-(k-1))} \\
&= \left(\frac{M-m}{M}\right)\left(\frac{M-1-m}{M-1}\right)\cdots\left(\frac{M-(k-1)-m}{M-(k-1)}\right) \\
&= \left(1-\frac{m}{M}\right)\left(1-\frac{m}{M-1}\right)\cdots\left(1-\frac{m}{M-(k-1)}\right) \\
&\approx (1-q)^{k(v_i)} \text{ for } k(v_i) \text{ relatively small compared to } m \text{ and } M.
\end{aligned}$$

This is simply the probability that a node of degree $k(v_i)$ loses all edges during subsampling $q^0(1-q)^k$ and thus $P(\text{not detecting } v_i) \approx (1-q)^{k(v_i)}$. Thus,

$$\hat{N} = \sum_{v_i \in V^*} \frac{1}{\pi_i} \tag{4.25}$$

$$= \sum_{v_i \in V^*} \frac{1}{1-(1-q)^{k(v_i)}} \tag{4.26}$$

$$\tag{4.27}$$

We apply these methods to our simulated and empirical networks.

Once $\hat{N}$ and $\hat{M}$ have been computed, the average degree is simply $\hat{k}_{\text{avg}} = \frac{2\hat{M}}{\hat{N}}$. The max degree scales roughly linearly for preferential attachment models and many of the empirical networks, however networks with high proportion of regular lattice structure (e.g.,

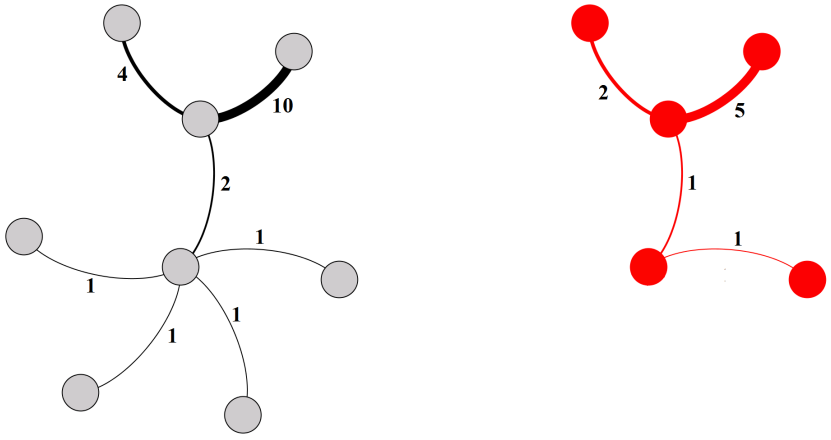
Small world and Range dependent) scale sublinearly. Clustering scales approximately as $\hat{C} = \frac{c}{q}$ and the giant component shows a critical threshold which varies according to network type and average degree. The relative error of our predictors are summarized in Tables 4.A11- 4.A15. Small world and Range dependent) scale sublinearly. Clustering scales approximately as $\hat{C} = \frac{c}{q}$ and the giant component shows a critical threshold which varies according to network type and average degree. The relative error of our predictors are summarized in Tables 4.A11- 4.A15.

To test the goodness of fit for the estimated degree distribution and the true $Pr(k)$, we again compute $D = \max |F_{i,true} - F_{i,predicted}|$, two sample Kolmogorov-Smirnov test statistic (Fig. 4.A18). This figure shows that reasonable results are achieved when $q > 50\%$, a noticeable increase in the percent of network knowledge needed, as compared to other sampling strategies (sampling by nodes and failing links).

4.4.4 Sampling by interactions

Lastly, we consider the case of sampling by interactions in the special case of a weighted network (Fig. 4.4). In this case, we begin with $G = (V, E)$, where E is a set of undirected[7] edges, e_j, with weight $w(e_j)$. The weight on an edge represents the number of interactions between two vertices and an alternative representation is simply a network with multiple edge between two such vertices, one for each interaction. A subnetwork generated by $q \sum_{e_j \in E} w(e_j)$ sampled interactions is simply a sampled collection of multi-edges and the nodes incident to these edges (e.g., the subnetwork generated by links with nonzero weight and nodes incident to those edges).

[7]We will treat the directed case as a special case at the end of this section.



(a) Weighted network (b) Weighted subnetwork

Figure 4.4: Subsampling by interactions in a weighted network. (a.) An unsampled weighted network consists of nodes, links and weights representing the number of interactions represented by the link. (b.) Sampling by interacting produces a subsample whereby links are included in the subsample only if at least one interaction has been sampled. The subnetwork is the induced subgraph on these links with $w_i \geq 1$.

To consider how the number of nodes scales, we consider a similar formulation as discussed in the previous section for the probability that a given node is selected when sampling by links, however instead of the degree of a node, $k(v_i)$, we are now interested in the strength of a node. The strength of a node is given by $s(v_i) = \sum_{e_j \in \mathcal{N}(v_i)} w(e_j)$, where $\mathcal{N}(v_i)$ denotes the neighborhood of vertex v_i [46]. Let $L = \sum_{e_j inE} w(e_j)$ represent network load and $\ell = qL$, the number of sampled interactions. If the number of interactions which are not sampled $(L - \ell)$ is less than the strength of a node $(s(v_i))$, then we can be certain that node v_i will be detected in sampling.

On the other hand, if $L - \ell \geq s(v_i)$, then observe that there are at most $\binom{L-s(v_i)}{\ell}$ ways[8] of choosing ℓ interactions from the $L - s(v_i)$ interactions not involving node v_i and there are at most $\binom{L}{\ell}$ total ways of choosing ℓ (distinct, labeled) interactions from all L. Letting

[8] As an upper bound, we assume that the $L - s(v_i)$ interactions are distributed over $L - s(v_i)$ edges (weight of 1 on each edge) which maximizes the number of ways these could be chosen.

$\mu(i)$ represent the probability that v_i is sampled, we have

$$\begin{aligned}\mu_i &= 1 - P(\text{ no interaction incident to } v_i \text{ is sampled}) \\ &= \begin{cases} 1 - \frac{\binom{L-s(v_i)}{\ell}}{\binom{L}{\ell}}, & \text{if } \ell \leq L - s(v_i) \\ 1, & \text{if } \ell > L - s(v_i). \end{cases}\end{aligned}$$

Thus, our Horvitz-Thompson estimator is,

$$\hat{N} = \sum_{v_i \in V^*} \frac{1}{\mu_i}, \tag{4.28}$$

where $\mu_i = P(v_i \text{ is sampled})$. This can be well approximated by

$$\mu_i = 1 - (1 - q)^{s(v_i)}. \tag{4.29}$$

It should be noted that the strength of a node is the predicted strength of a node and thus effort must be made to predict the node strength distribution in the same spirit as was previously done for the degree distribution. To predict the node strength distribution, we modify Equation 4.17 and predict that an observed node of strength s to be of strength $\frac{s}{q}$ in the true network. Applying this to corrector to our subsampled weighted networks, we find low relative error in the predicted number of nodes for most networks (Tables 4.A16 and 4.A17). An exception to this is Case I (Erdös-Rényi) for $q < 0.55$. In this case, we are essentially predicting a node strength by $\frac{s}{q} \geq 2$ and yet in this case, the true network is unweighted (e.g., $w(e_j) = 1, \forall e_j \in E$). If there is knowledge that the network is unweighted, this example shows that the techniques from Section 4.4.3 will yield much better results.

We now consider how the number of edges in the subnetwork scales with the proportion of sampled interactions. The probability of selecting an edge $e_j \in E$ is equal to 1-$P($ not selecting edge $e_j)$. Notice that when the $\ell > L - w(e_j)$, the edge e_j is certain to be included in the subsample. When $\ell \leq L - w(e_j)$, the probability of not selecting edge e_j is simply the number of ways of selecting the $L - w(e_j)$ interactions ℓ at a time, which are not on edge e_j divided by the number of ways of selecting ℓ weights from L.

$$\begin{aligned} P(e_j \text{ is sampled}) &= 1 - P(\text{ no interaction along } e_j \text{ is sampled}) \\ &= \begin{cases} 1 - \frac{\binom{L-w(e_j)}{\ell}}{\binom{L}{\ell}}, & \text{if } \ell \leq L - w(e_j) \\ 1, & \text{if } \ell > L - w(e_j). \end{cases} \end{aligned}$$

Thus, our Horvitz-Thompson estimator is,

$$\hat{M} = \sum_{e_j \in E^*} \frac{1}{\lambda_j}, \tag{4.30}$$

where $\lambda_j = P(e_j \text{ is observed})$, which is well approximated by

$$\lambda_j = 1 - (1 - q)^{w(e_j)}. \tag{4.31}$$

Again, we must have knowledge of the edge weights, or be able to predict them with reasonable accuracy. To do this, we predict an edge of weight $w(e_j)$ in the subnetwork to be of edge weigh $\frac{w(e_j)}{q}$ in the true network.

As the weights on edges tends to 1 (the unweighted network case), we retrieve our result for how edges scale when links (syn. with weights in the case where $w_i = 1$) are sampled.

$$\begin{aligned}
\lim_{w(e_j)\to 1} P(e_j \text{ is observed}) &= \lim_{w(e_j)\to 1} 1 - P(w(e_j)) \\
&= \lim_{w(e_j)\to 1} 1 - \frac{\binom{L-w(e_i)}{\ell}}{\binom{L}{\ell}} \\
&= 1 - \frac{\binom{M-1}{m}}{\binom{M}{m}} \\
&= 1 - \frac{M-m}{M} \\
&= \frac{m}{M} \\
&= q,
\end{aligned}$$

where q is the proportion of sampled links. Thus, when the weights on edges tends to 1, our Horvitz-Thompson estimator is

$$\begin{aligned}
\hat{M} &= \sum_{e_j \in E^*} \frac{1}{\lambda_j}, \\
&= \frac{m}{q},
\end{aligned}$$

which recovers our previous result for scaling of edges when sampling by links. The relative error incurred for the predicted number of edges is presented in Tables 4.A18 and 4.A19.

Having found suitable predictors for N and M, the average degree may be predicted by,

$$\hat{k}_{\text{avg}} = \frac{2\hat{M}}{N}.$$

Applying these scaling techniques, we obtain reasonably low error for both networks in both experiments 1 and 2 (Tables 4.A20- 4.A21).

To estimate $k_{\max}$, we recognize that the observed max degree will need to be scaled by roughly the proportion of missing edges. Using $\frac{\hat{M}}{m}$ as our scaling factor, we find relatively high error for both networks (Tables 4.A22- 4.A23) and this is due to errors in the $\hat{M}$ may be hindering $\hat{k}_{\max}$.

Table 4.2: Summary of scaling techniques.

	Sampled nodes	Failed links	Sampled links	Sampled interactions
$\hat{N}$	$\frac{n}{q}$	n	$\sum_{v_i \in V^*} \frac{1}{1-(1-q)^{d(v_i)}}$	$\sum_{v_i \in V^*} \frac{1}{1-(1-q)^{s(v_i)}}$
$\hat{M}$	$\frac{m}{q^2}$	$\frac{m}{q}$	$\frac{m}{q}$	$\sum_{e_i \in E^*} \frac{1}{1-(1-q)^{w(e_i)}}$
$\hat{k}_{\text{avg}}^{obs}$	$\frac{k_{\text{avg}}^{obs}}{q}$	$\frac{k_{\text{avg}}^{obs}}{q}$	$\frac{2\hat{M}}{\hat{N}}$	$\frac{2\hat{M}}{\hat{N}}$
$\hat{C}$	C	qC	$\frac{C}{q}$	–
$\hat{k}_{\max}$	$\frac{k_{\max}^{obs}}{q}$	$\frac{k_{\max}^{obs}}{q}$	$\frac{k_{\max}^{obs}}{q}$	$\frac{\hat{M}}{m} \cdot k_{\max}^{obs}$

4.5 Estimating the size of the Twitter interactome

In this section we consider the weighted, directed network of replies whereby a link from node v_i to node v_j represents the existence of at least one reply directed from v_i to v_j and

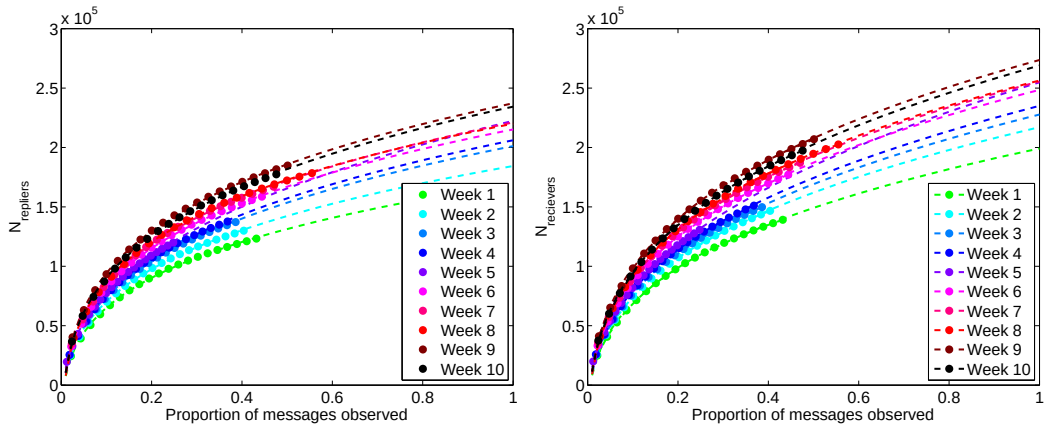
the weight on this edge represents the number of messages sent in the time period under consideration. We apply our methods to reply networks constructed from tweets gathered during the ten week period from September 9, 2008 to November 17, 2008, a period for which we have a substantially higher percentage of all authored messages.

For each of these weeks, we receive between 20-55% of all messages posted on Twitter and similarly believe that we receive approximately 20-55% of all replies posted in this period (Table 4.A24). We apply our previously developed methods to estimate the number of nodes, edges, strengths on these edges, average degree, max degree and distribution of node strength. To help validate our predictions, we also predict the number of nodes, edges, average degree and max degree by performing 100 sampling experiments in which a proportion q of the observed messages used for subnetwork construction. These sampling experiments essentially "hide" some of the messages from our view and thus allow us to consider how further subsampling impacts the inferred networks statistics. Curve fitting over this region of q allows us to extrapolate the network statistic to a predicted value over increased percentages of observed messages. We use this to validate with our estimated parameter using the methods from the previous section.

4.5.1 Number of nodes

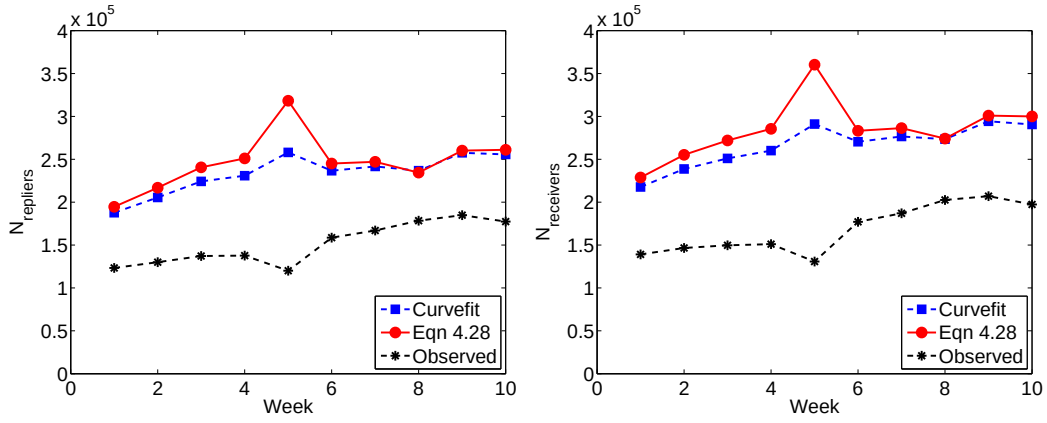
Since our reply networks are directed, we consider both the number of nodes which make a reply (N_{replier}) and the number of nodes which receive a reply (N_{receiver}). As expected from our previous discussion, the number of nodes scales nonlinearly with the proportion of observed messages (Fig. 4.5). We fit models of the form $N = ax^b$ to observed data and in doing so find an excellent fit ($R^2 \approx 0.99$) for all weeks over the subsampled region (Fig. 4.5). Extrapolating these fitted models to $q = 1$, we find excellent agreement with our

predicted number of nodes obtained from Equations 4.28 and 4.29. The predicted number of nodes agrees from both methods agree to within $\pm$ 5%.



(a) N_{repliers} (b) $N_{\text{receivers}}$

Figure 4.5: Number of nodes in Twitter reply subnetworks. (a.) The N_{repliers} is shown for Weeks 1 to 10, where each data point (dot) represents the average over 100 simulated subsampling experiments. The dashed line represents the best fitting model of the form $N_{\text{repliers}} = ax^b$ to the observed data. We extrapolate this model to predict N_{repliers}. (b.) The same as panel (a.), except for $N_{\text{receivers}}$.

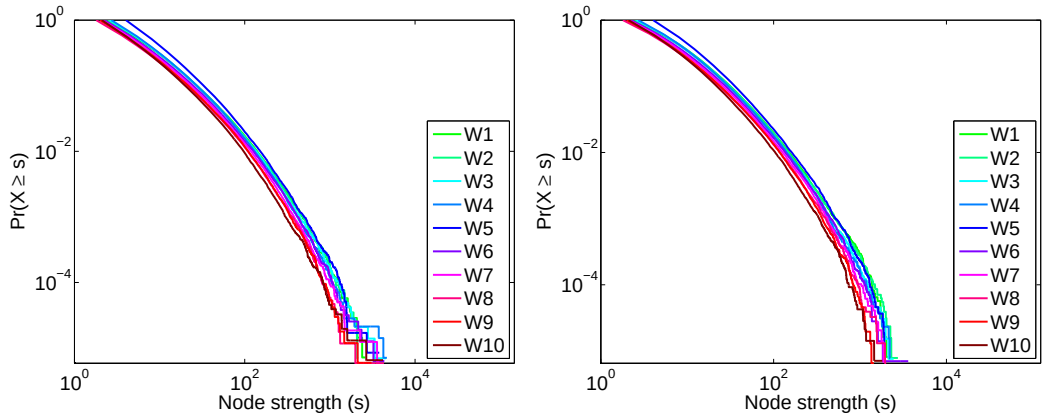


(a) N_{repliers} (b) $N_{\text{receivers}}$

Figure 4.6: Predicted number of nodes in Twitter reply networks. The relatively low proportion of messages received for Week 5 ($< 25\%$) may be creating greater inaccuracies in the predictors for that week.

4.5.2 Strength of nodes

Figure 4.7 depictes a log-log plot of the predicted node strength distribution. This plot reveals that there are fewer nodes in the high strength region than would be expected in a scale-free node strength distribution. Figure 4.8 reveals that low degree nodes dominate the dataset and that many of these low degree nodes often have low average edge weight ($w_{\text{avg}} \approx 1.5$). We also find a peak in the average weight per edge as a function of degree around $k \approx 10^2$ (Dunbar's number [47]). This peak is large for more pronounced for out-going edges and may suggest that beyond this value, a limiting factor may prevent increases in the weight per edge, a result also noted by Gonçalves et al. (31).



(a) In-coming (b) Out-going

Figure 4.7: Predicted $Pr(s)$ for Twitter reply networks. (a.) The node strength distribution for in-coming interactions. (b.) The node strength distribution out-going interactions. In both cases, the distribution is heavy tailed, but falls off faster than would be expected in a scale-free distribution.

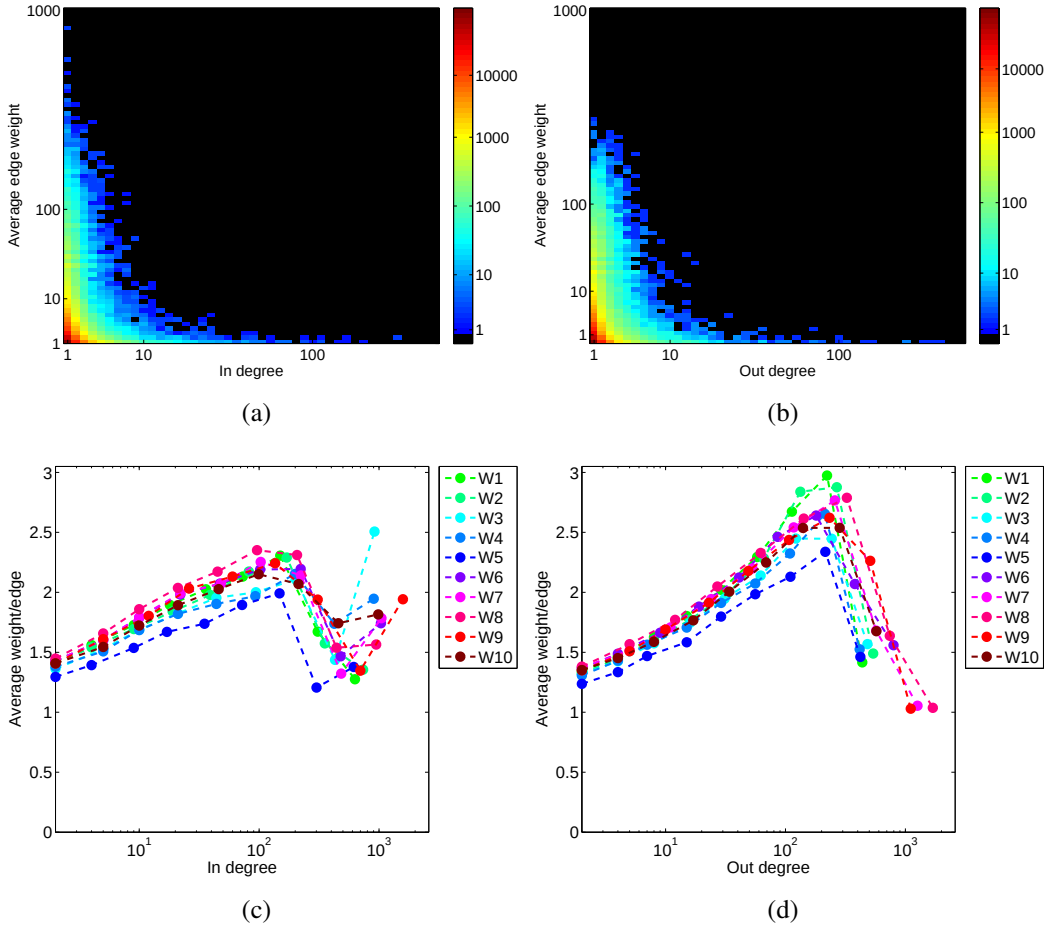
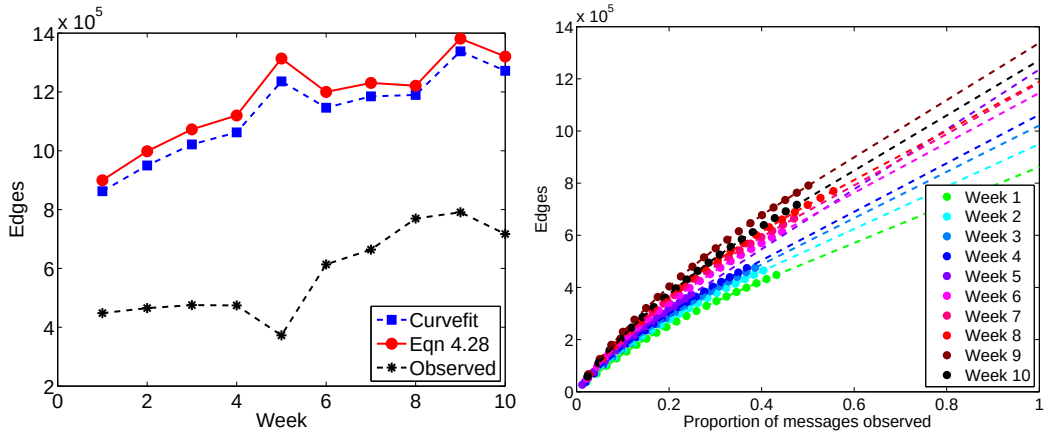


Figure 4.8: In, Out-degree vs. Average edge weight for Twitter reply networks. (a.) The average in-coming edge weight for each node of degree k is depicted in a logarithmically binned heatmap. (b.) The same as (a), except for out-going edges. (c.) The average weight per edge for in-coming edges as a function of k_{in} shows a gradual increase to $k_{in} \approx 10^2$ with a peak of approximately 2.2 interactions per edge. (d.) The average weight per edge for out-going edges as a function of k_{out} shows a gradual increase to $k_{out} \approx 10^2$ with a peak of between 2.5 and 3 interactions per edge.

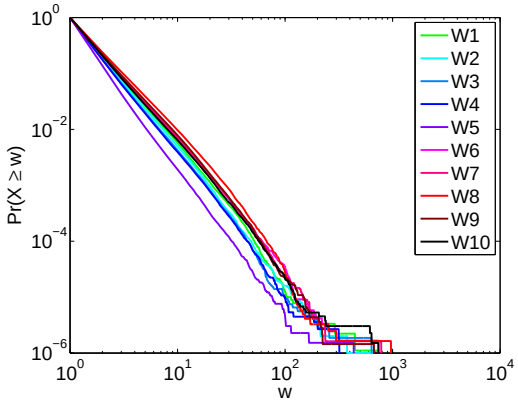
4.5.3 Number of edges

The number of edges can be predicted using Equations 4.30 and 4.31. We present our results in Figure 4.9. In all cases, the number of edges increases throughout the period of the study. Figure 4.10 depicts the predicted edge weight and degree distributions. The edge weight distribution shows that very few ($< .001\%$) edges have an edge weight greater than 10^2. The degree distribution of the observed subnetwork can be rescaled by reassigning nodes of degree k, to nodes of degree $\frac{\hat{M}}{m}k$. Figure 4.10bc demonstrates a slightly heavier tail in the in-degree distribution as compared to the out-degree distribution. The degree distribution reveals that fewer than .01% of the nodes have more than 10^2 distinct neighbors. This value is approximately Dunbar's number, a value suggested to be the upper limit on the number of active social contacts for humans (47).

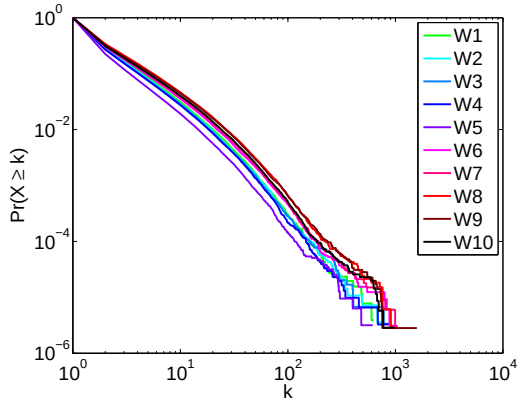


(a) (b)

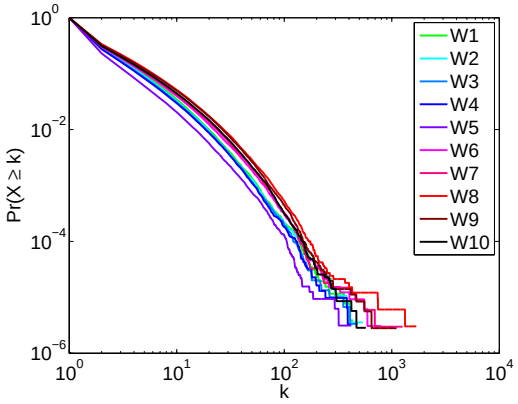
Figure 4.9: Predicted number of edges in Twitter reply networks. (a.) A small proportion of observed messages for Week 5 ($< 25\%$) may explain the spike in the estimated number of edges for that week. (b.) Each data point represents the number of directed edges observed, averaged over 100 simulated subsampling experiments. The dashed line extrapolates the predicted number of edges for greater proportions of sampled data.



(a) Edge weights



(b) $Pr(k_{in})$

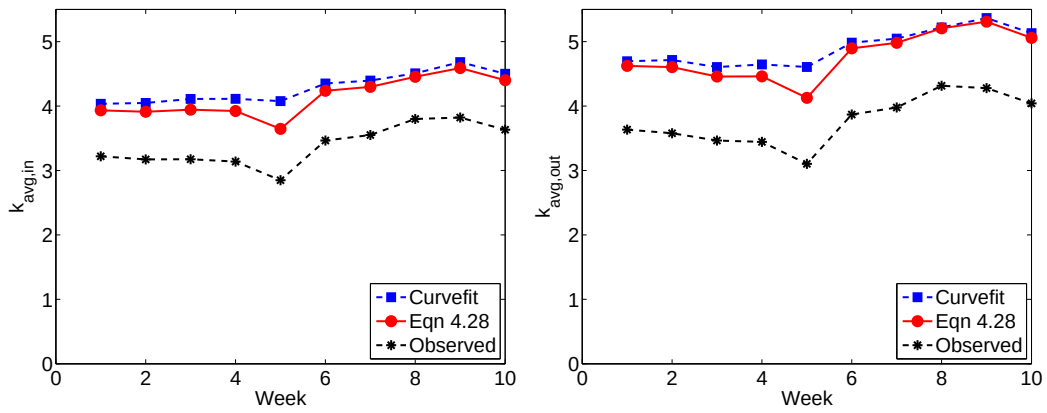


(c) $Pr(k_{out})$

Figure 4.10: Predicted edge weight and degree distributions for Twitter reply networks. (a.) The predicted edge weight distribution. (b.) Predicted $Pr(k_{in})$ and (c.) $Pr(k_{out})$ for Twitter reply networks.

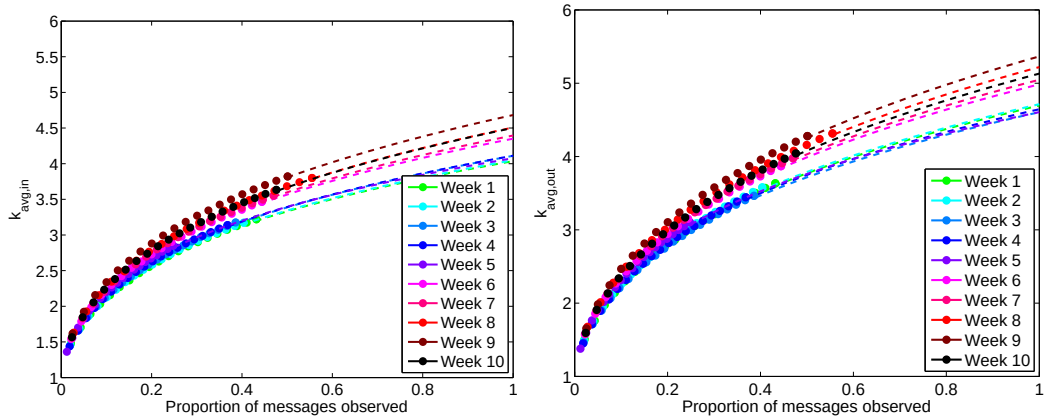
4.5.4 Average degree

Once the number of nodes and edges have been predicted for the network, we may simply compute the average degree as $\hat{k}_{avg,in} = \frac{\hat{M}}{\hat{N}_{receivers}}$ and $\hat{k}_{avg,out} = \frac{\hat{M}}{\hat{N}_{repliers}}$. Upon doing so, we find that the average degree for Twitter reply networks by between 4 and 5 (Fig. 4.11). We find that the average in-degree is less than the average out-degree.



(a) $k_{\text{avg,in}}$ (b) $k_{\text{avg,out}}$

Figure 4.11: Predicted $k_{avg,in}$ and $k_{avg,out}$ in Twitter reply networks.



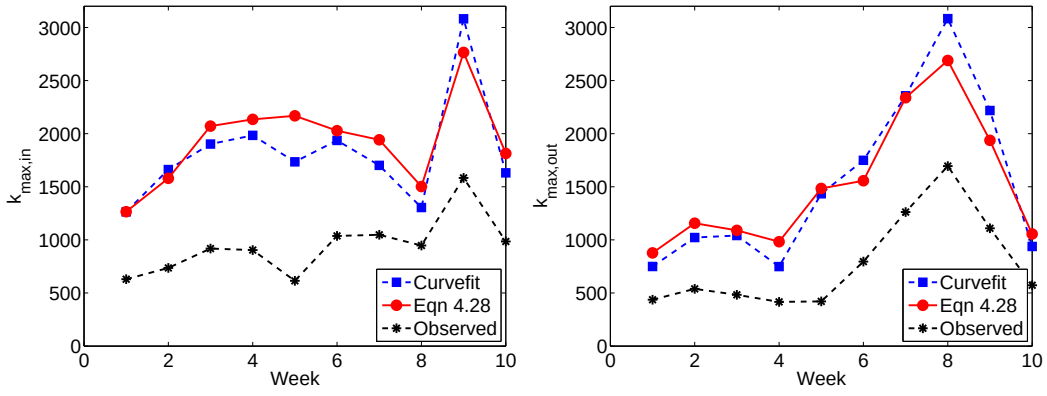
(a) $k_{\text{avg,in}}$ (b) $k_{\text{avg,out}}$

Figure 4.12: $k_{\text{avg,in}}$ and $k_{\text{avg,in}}$ for Twitter reply networks. Each data point represents the observed average in- and out-degree, averaged over 100 simulated subsampling experiments. The dashed line extrapolates the predicted number of edges for greater proportions of sampled data.

4.5.5 Maximum degree

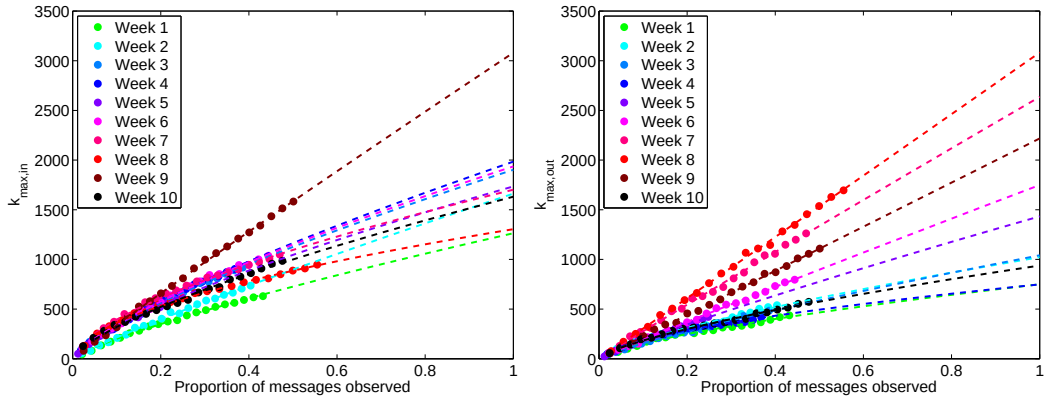
The maximum degree simply scales in proportion to the probability of edge inclusion. Since the probability of edge inclusion is no longer q, as in the case of sampling by links,

we may approximate the probability of edge inclusion by $\frac{m}{\hat{M}}$ and thus $\hat{k}_{\max} = \frac{\hat{M}}{m} k_{\max}^{obs}$ as mentioned in the previous section. The predicted maximum degree for Twitter reply networks is shown in Figures 4.13 and 4.14.



(a) $k_{\max,in}$ (b) $k_{\max,out}$

Figure 4.13: Predicted $k_{\max,in}$ and $k_{\max,out}$ in Twitter reply networks.



(a) $k_{\max,\text{in}}$ (b) $k_{\max,\text{out}}$

Figure 4.14: $k_{\max,in}$ and $k_{\max,in}$ for Twitter reply networks. Each data point represents the observed maximum in- and out-degree, averaged over 100 simulated subsampling experiments. The dashed line extrapolates the predicted number of edges for greater proportions of sampled data.

4.6 Discussion

Network measures derived from empirical networks will often be poor plug-in estimators of the true underlying network structure of the system. We have explored four sampling regimes: (1) subnetworks induced on randomly sampled nodes, (2) subnetworks obtained when all nodes are known and some links fail or are hidden, (3) subnetworks generated from randomly sampled links and (4) weighted subnetworks generated by randomly sampled interactions. We have described how network statistics scale under these regimes via sampling experiments on simulated and empirical networks. Our paper advances an understanding of how network statistics scale, and more importantly how to correct for missing data when the proportion of missing nodes, links or interactions is known.

A major obstacle to generating scaling techniques for subnetworks generated by sampled links or interactions has previously been the lack of a practical method for estimating the true degree distribution or node strength distribution. Problematically, the random selection of links creates a biased sample of nodes whereby hubs are more likely to be detected and nodes of small degree are more likely to go undetected. Although scaling methods have been suggested, they are based on knowledge of (or a reasonable estimate of) the degree or node strength distribution (3). In this paper, we have overcome this obstacle by our proposed scaling techniques for the degree distribution and apply this to several simulated and empirically derived networks with reasonably good results.

Very few studies have addressed the missing data problem in empirically studied networks, such as those constructed from tweets. An exception is work by Morstatter et al. (2013) who compares network statistics for the current Twitter streaming API ($\approx 1\%$ of all

tweets) to the full Firehose (100% of all tweets), however no methods for scaling from data collected via the API are suggested.

We conclude our work by applying our derived scaling methods to Twitter reply networks. Our work supports Dunbar's hypothesis which suggests that individuals maintain an upper limit of roughly 100-150 contacts each week (47). Further evidence for this hypothesis comes from previous work in link prediction effort which detects the Resource Allocation index to be one that often evolves to have a large, positive weight - thus contributing heavily (and positively) in the prediction of new links (49). This index considers the amount of time and attention one individual has as a "social resource" to spend in the social network and assumes that each node will distribute its resource equally among all neighbors. Although the presence of hubs is suggestive of preferential attachment, it is clear that the constraints of time and attention limit truly scale-free behavior in weekly Twitter reply networks. We find that the number of individuals who make replies is less than the number of individuals who receive replies.

One limitation of our work is that our scaling methods are based upon the assumption that q is known, while in practice this need not be the case. In cases where one may establish an upper and lower bound for q, our methods could be used to help establish bounds for the predicted network measures. In some cases, particularly when sampling by links or interactions, small changes in q may have relatively little impact on the predicted parameters, especially for large q. Future work that seeks to classify subnetworks by network class based on signature subsampling properties may also prove to be fruitful. With some knowledge of network class or generative model, methods for estimating q may be possible. Additionally, efforts to predict structural holes in networks from localized information may also greatly advance the field (50).

To our knowledge, this is the first attempt provide scaling methods for $k_{\max}$. While our scaling techniques for predicting $k_{\max}$ perform well for several networks, they did not perform as well on simulated networks with a regularized structure.[9] Future work which detect and accounts motif distributions may improve upon our efforts here.

With an increased interest in large, networked datasets, we hope that continued efforts aid in the understanding of how subsampled network data can be used to infer properties of the true underlying system. Our methods advance the field in this direction, not only adding to the body of literature surrounding sampling issues and Twitter's API (2), but also to the growing body of literature on incomplete network data.

4.7 Acknowledgments

The authors acknowledge the Vermont Advanced Computing Core which is supported by NASA (NNX-08AO96G) at the University of Vermont for Providing High Performance Computing resources that have contributed to the research results reported within this paper. CAB and PSD were funded by an NSF CAREER Award to PSD (# 0846668). CMD and PSD were funded by a grant from the MITRE Corporation.

4.8 References

[1] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 631–636, New York, NY, USA, 2006. ACM.

[9] Our rewiring probability for the simulated Small world networks was quite low, with $p = 0.1$. Our methods perform well on other networks which are known to exhibit to Small world structure, such as our empirical networks Powergrid and *C. elegans*.

[2] Fred Morstatter, Jurgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? Comparing data from Twitters streaming API with Twitters firehose. *Proceedings of ICWSM*, 2013.

[3] Eric D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. New York, NY: Springer Publishing Company, Inc., 1st edition, 2009.

[4] E. Costenbader and T. W. Valente. The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307, 2003.

[5] J. D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23:839–944, 2005.

[6] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4221–4224, 2005.

[7] G. Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3):247–268, 2006.

[8] C. Wiuf and M. P. H Stumpf. Binomial subsampling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 462(2068):1181–1195, 2006.

[9] M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.

[10] T.L. Frantz, M. Cataldo, and K.M. Carley. Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328, 2009.

[11] S. Martin, R. D. Carr, and J.-L. Faulon. Random removal of edges from scale free graphs. *Physica A: Statistical Mechanics and its Applications*, 371(2):870 – 876, 2006.

[12] E. de Silva, T. Thorne, P. Ingram, I. Agrafioti, J. Swire, C. Wiuf, and M. Stumpf. The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biology*, 4(1):39, 2006.

[13] A. Lakhina, J. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurements. In *Proceedings of IEEE Infocom*, April 2003.

[14] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(1):016102, 2006.

[15] O. Frank and T. Snijders. Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10:53–53, 1994.

[16] P. Biernacki and D. Waldorf. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods and Research*, 10(2):141–163, 1981.

[17] D. D. Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, pages 174–199, 1997.

[18] D. D. Heckathorn, S. Semaan, R. S. Broadhead, and J. J. Hughes. Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18–25. *AIDS and Behavior*, 6(1):55–67, 2002.

[19] S. Semaan, J. Lauby, and J. Liebman. Street and network sampling in evaluation studies of hiv risk-reduction interventions. *AIDs Review*, 4(4):213–23, 2002.

[20] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy*, 98:404–409, 2001.

[21] Ove Frank. Sampling and estimation in large social networks. *Social Networks*, 1(1):91 – 101, 19781979.

[22] P. Erdös and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 5:17–61, 1960.

[23] Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

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

[25] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[26] G. U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87, 1925.

[27] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[28] M. P. H. Stumpf and C. Wiuf. Sampling properties of random graphs: the degree distribution. *Physical Review E*, 72(3):036118, 2005.

[29] Ove Frank. Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference*, 4(1):45 – 50, 1980.

[30] J. Platig, M. Girvan, and E. Ott. Robustness of network measures to link errors. *Bulletin of the American Physical Society*, 58, 2013.

[31] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, 85(21):4626, 2000.

[32] M. Stumpf, P. Ingram, I. Nouvel, and C. Wiuf. Statistical model selection methods applied to biological networks. *Transactions on Computational Systems Biology III*, pages 65–77, 2005.

[33] C. A. Bliss, I. M. Kloumann, K. D. Harris, C. M. Danforth, and P. S. Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, 3(5):388 – 397, 2012.

[34] D. D. S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

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

[36] P. Grindrod. Range-dependent random graphs and their application to modeling large small-world Proteome datasets. *Physical Review E*, 66:066702, 2002.

[37] A. Taylor and D. J. Higham. CONTEST: A controllable test matrix toolbox for MATLAB. *ACM Transactions on Mathematical Software*, 35:26:1–26:17, February 2009.

[38] J.G. White, E. Southgate, J.N. Thompson, and S. Brenner. The structure of the nervous system of the nematode *C. Elegans*. *Philosophical Transactions of the Royal Society of London*, 314:1–340, 1986.

[39] O. Woolley-Meza, D. Grady, C. Thiemann, J. P. Bagrow, and D. Brockmann. Eyjafjallajökull and 9/11: The impact of large-scale disasters on worldwide mobility. *PloS one*, 8(8):e69829, 2013.

[40] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, pages 452–473, 1977.

[41] D. Lusseau, K. Schneider, O. Boisseau, P. Haases, E. Slooten, and S. Dawson. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54:396–405, 2003.

[42]

[43] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han. Attack vulnerability of complex networks. *Physical Review E*, 65(5):056109, 2002.

[44] A. Barrat, M. Barthlemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge University Press, 2008.

[45] M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41(2):255–258, 2004.

[46] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

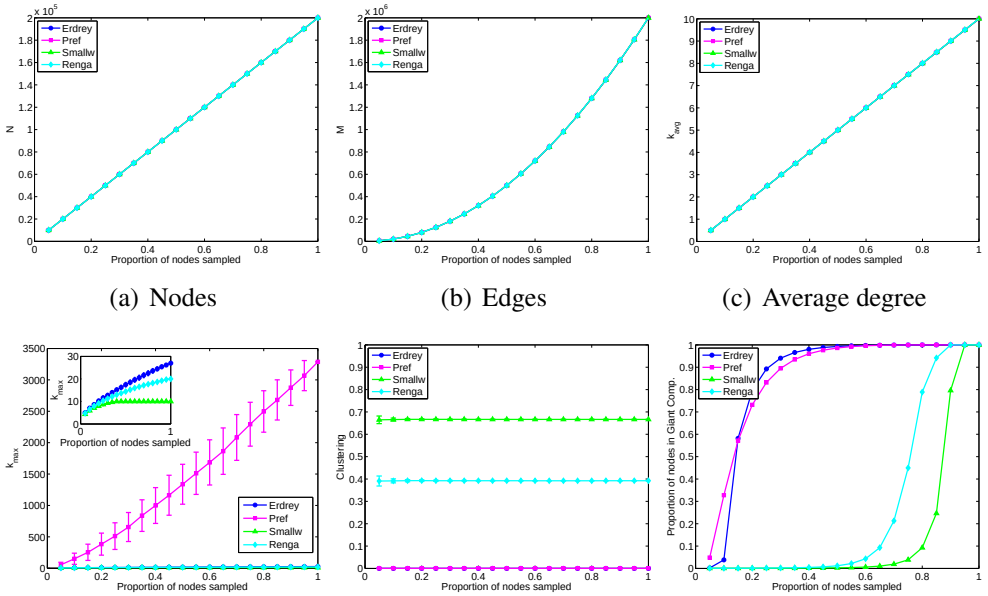
[47] R. I. M. Dunbar. Neocortex size and group size in primates: A test of the hypothesis. *Journal of Human Evolution*, 28(3):287 – 296, 1995.

[48] B. Gonçalves, N. Perra, and A. Vespignani. Modeling users' activity on Twitter networks: Validation of Dunbar's Number. *PLoS one*, 6, 08 2011.

[49] C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds. An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science*, 2014.

[50] J. P. Bagrow, S. Desu, M. R. Frank, N. Manukyan, L. Mitchell, A. Reagan, E. E. Bloedorn, L. B. Booker, L. K. Branting, M. J. Smith, B. F. Tivnan, C. M. Danforth, P. S. Dodds, and J. C. Bongard. Shadow networks: Discovering hidden nodes with models of information flow. *arXiv preprint, arXiv:1312.6122*, 2013.

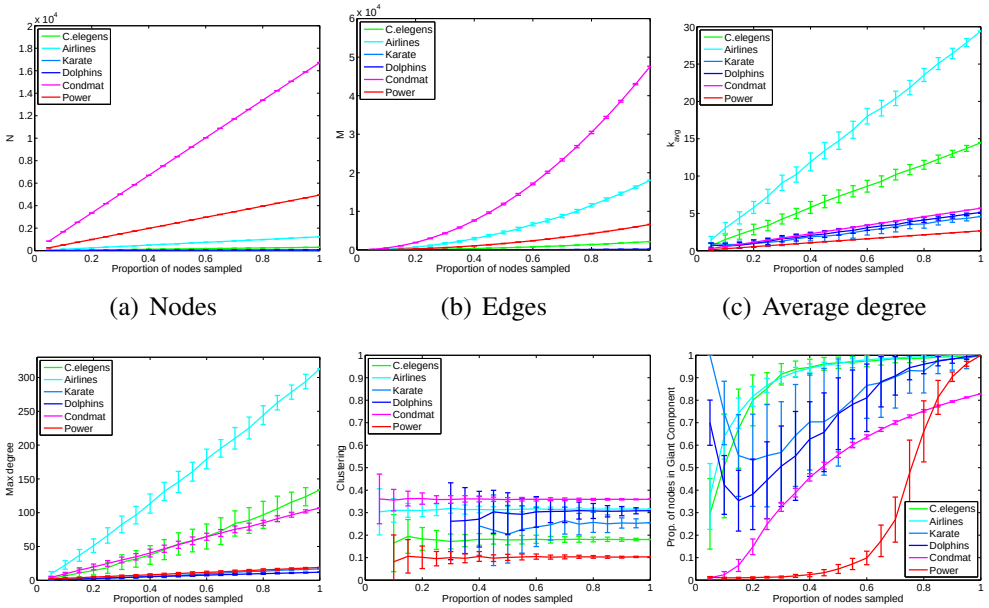
4.9 Appendix



(a) Nodes (b) Edges (c) Average degree

(d) Max degree (e) Clustering (f) Prop. of nodes in Giant Component

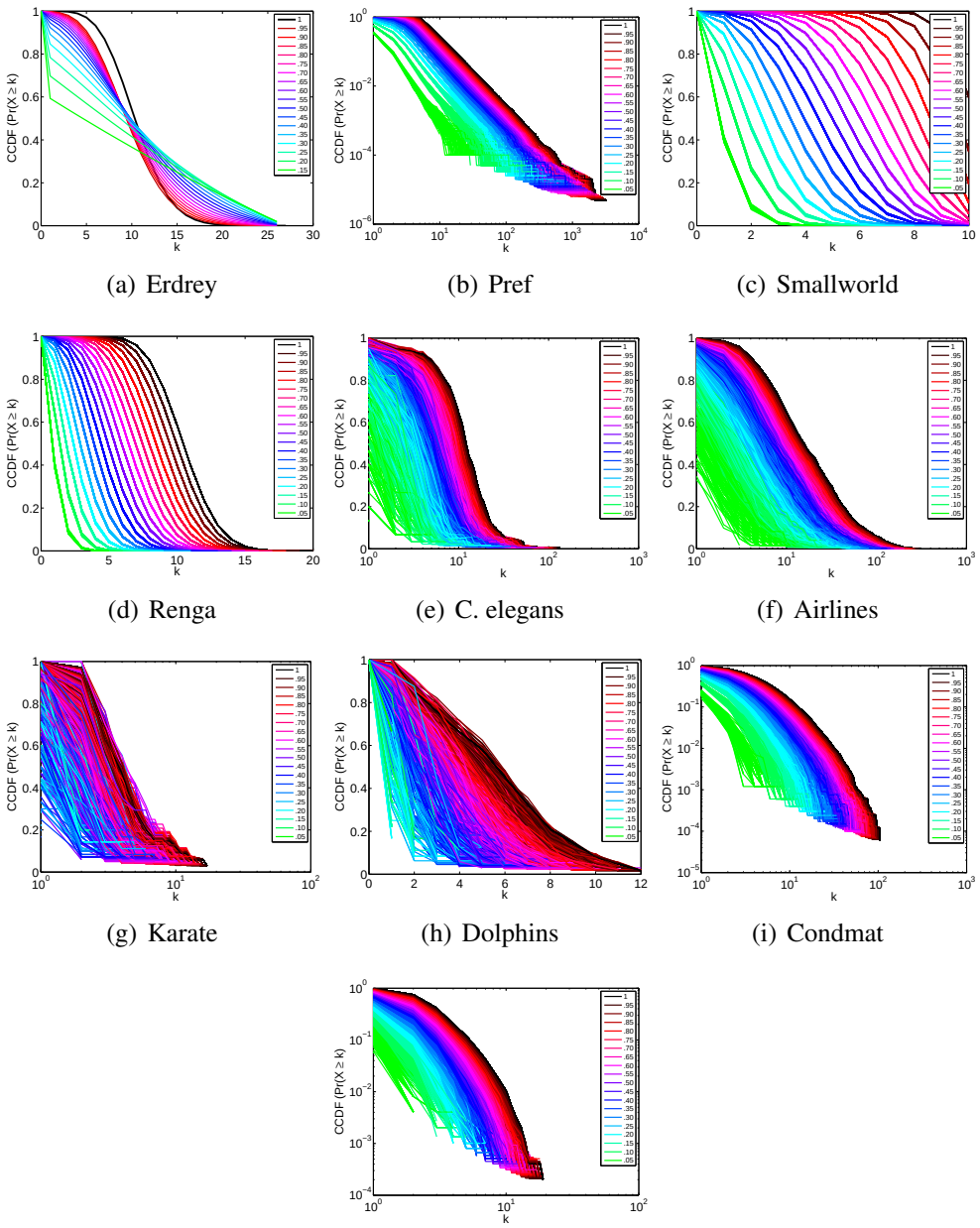
Figure 4.A1: Scaling of statistics for simulated subnetworks induced on sampled nodes. (a.) The number of nodes in a subnetwork sampled by nodes scales as $n = qN$ precisely because only qN nodes are selecting during subsampling. (b.) The number of edges scales as $m \approx M \cdot \frac{n(n-1)}{N(N-1)} \approx Mq^2$, for $n >> 1$ and $N >> 1$. (c.) The average degree scales linearly with the proportion of nodes subsampled. (d.) The scaling of the max degree is dependent on network type. For networks with few large hubs, $k^{obs}_{k_{\max}} \approx qk_{\max}$. For networks exhibiting a nontrivial number of nodes with degrees relatively close to $k_{\max}$, the max. degree scales nonlinearly. (e.) The clustering coefficient (20) shows little variation with respect to q as suggested by the analytical result from Frank (21). This suggests that $\hat{C} \approx C_{\text{obs}}$. (f.) The proportion of nodes in the giant component increases with the proportion of nodes sampled. For the random graphs (Erdrey and Pref) there is a critical point corresponding to the approximate sampling level corresponding to when $k^{obs}_{\text{avg}} > 1$. The thresholds for Small World and Range dependent networks are much higher due to the uniformity of the motif distribution in these networks. Markers indicates the mean over 100 simulations. Error bars showing one standard deviation are too small to see, except for (d.).



(a) Nodes (b) Edges (c) Average degree

(d) Max degree (e) Clustering (f) Prop. of nodes in Giant Component

Figure 4.A2: Scaling of statistics for empirical subnetworks induced on sampled nodes. (a.) The number of nodes scales as $n = qN$ precisely because only qN nodes are selecting during subsampling. (b.) The number of edges scales as $m \approx M \cdot \frac{n(n-1)}{N(N-1)} \approx Mq^2$, where q is the proportion of nodes subsampled. (c.) The average degree scales as $k_{\text{avg}}^{\text{obs}} \approx qk_{\text{avg}}^{\text{true}}$. (d.) The max degree scales roughly linearly as $k_{\text{max}}^{\text{obs}} \approx qk_{\text{max}}^{\text{true}}$. (e.) The clustering coefficient (20) shows little variation with respect to q as suggested by the analytical result from Frank (21), $\hat{C} \approx C_{\text{obs}}$. (f.) Large networks, such as the Powergrid and Condensed Matter author collaboration networks show the expected transition to the giant component as q increases corresponding to when $k_{\text{avg}}^{obs} > 1$. Smaller networks, such as the Karate club and Dolphin network show a high proportion of nodes in the giant component, for low q because the subnetwork generated for these levels of q contains fewer than 10 nodes (i.e., the network is degenerate).



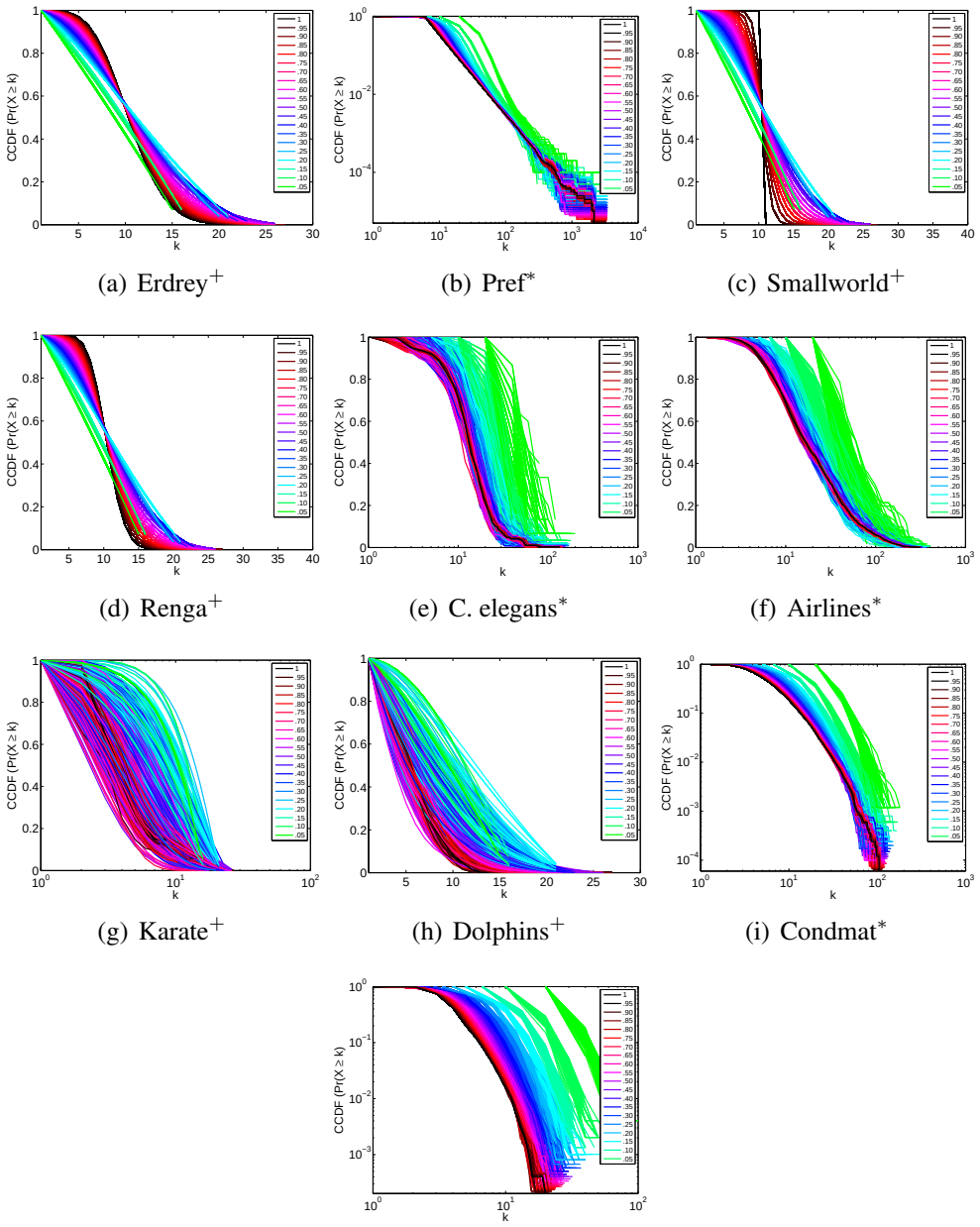
(a) Erdrey (b) Pref (c) Smallworld

(d) Renga (e) C. elegans (f) Airlines

(g) Karate (h) Dolphins (i) Condmat

(j) Powergrid

Figure 4.A3: CCDF distortion for subnetworks induced on sampled nodes. Subnetwork degree distributions do not capture the true degree distribution, especially for small q.



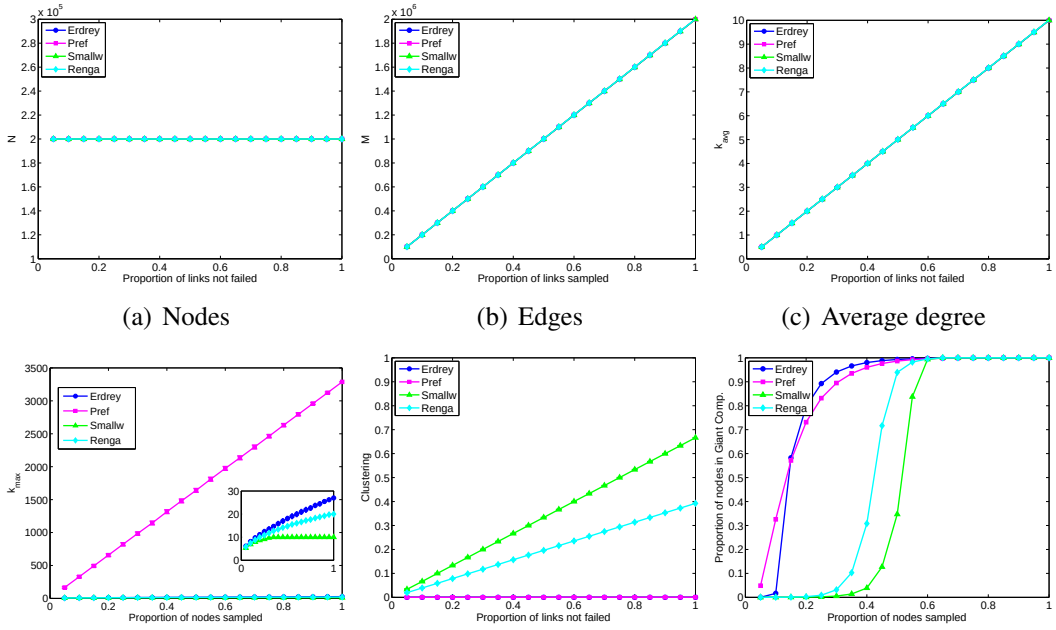
(a) Erdrey^{+} (b) Pref* (c) Smallworld^{+}

(d) Renga^{+} (e) C. elegans* (f) Airlines*

(g) Karate^{+} (h) Dolphins^{+} (i) Condmat*

(j) Powergrid*

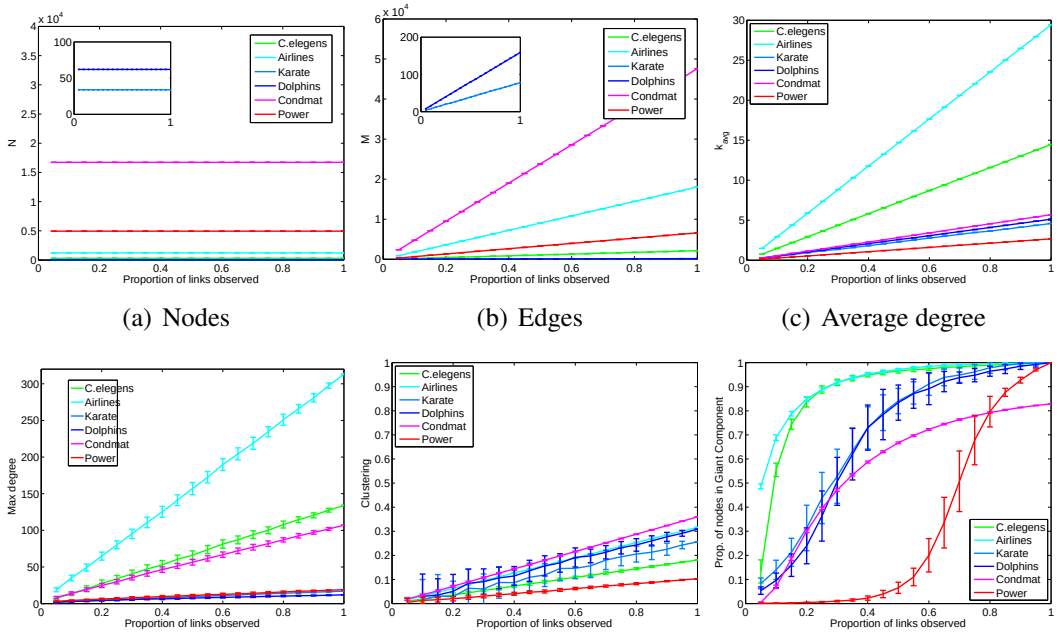
Figure 4.A4: Predicted CCDF from subnetworks induced on sampled nodes. The predicted CCDF shows relatively good agreement with the true CCDF for most networks. Karate club and Dolphins exhibit significant deviation, possible due to the small number of nodes in these networks. Networks designated with $^{+}$ utilized Equation 4.15 and those designated with with * utilized Equation 4.17.



(a) Nodes (b) Edges (c) Average degree

(d) Max degree (e) Clustering (f) Prop. of nodes in Giant Component

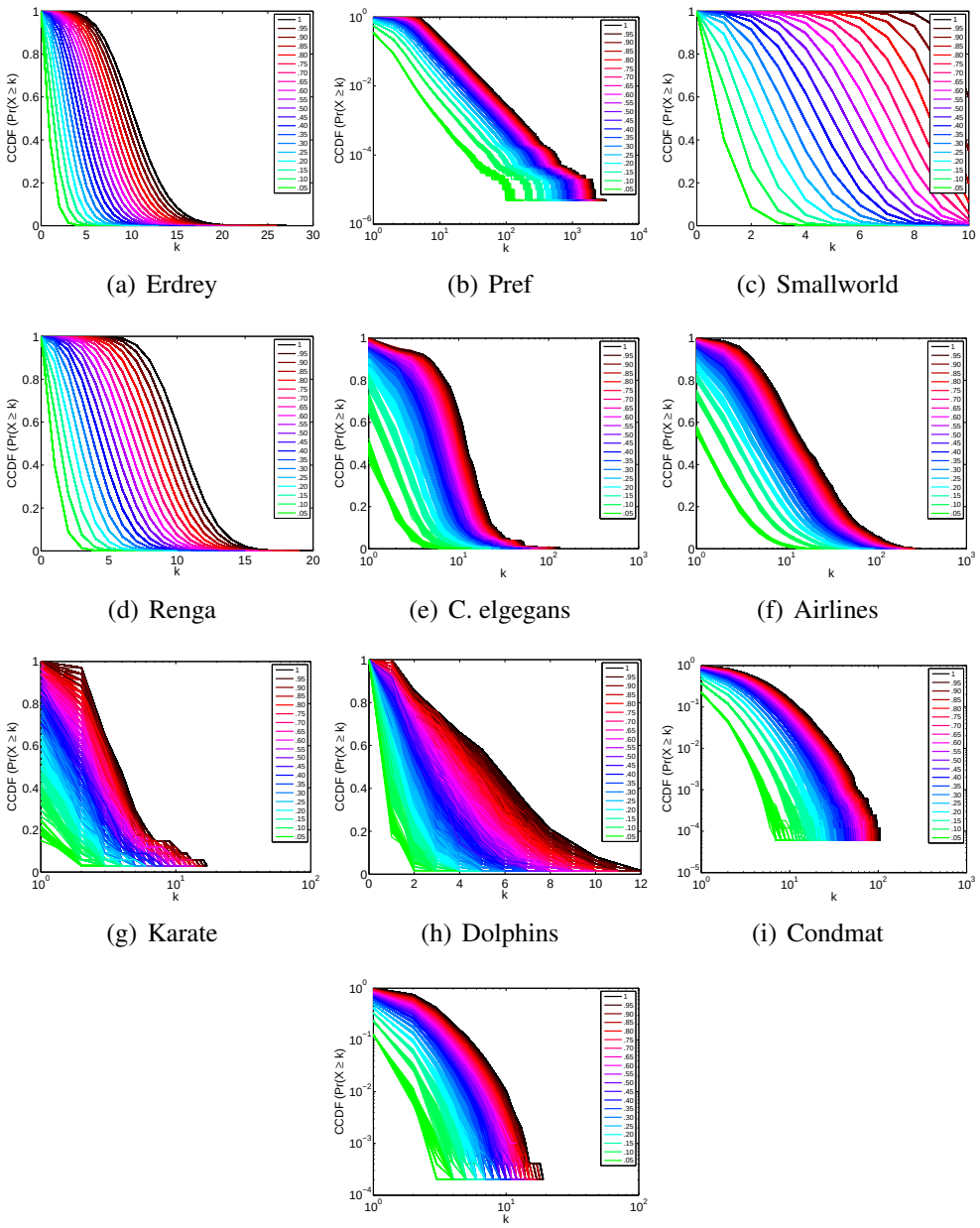
Figure 4.A5: Scaling of subnetwork statistics for simulated networks obtained by failing links. (a.) When all nodes are known q links are observed through sampling, the sample statistic for the number of nodes n equals the true number of nodes N. It should be noted, though, that some nodes of degree 0 may be observed and these are counted as nodes (not discarded). (b.) The number of edges scales linearly as $M_{\text{obs}} = qM$. (c.) The average degree scales linearly as $k_{\text{avg}}^{obs} = \frac{k_{\text{avg}}^{true}}{q}$. (d.) The max degree scales linearly for Pref, but nonlinearly for other networks which have several nodes with degree similar to $k_{\max}$. (e.) Clustering scales roughly linearly with q. (f.) The percolation threshold for random graphs (Erdös-Rényi and Preferential attachment) roughly corresponds to the q for which $k_{\text{avg}} \geq 1$. Smallworld and Renga show more fragility and have a threshold which is closer to $q \approx 0.4$.



(a) Nodes (b) Edges (c) Average degree

(d) Max degree (e) Clustering (f) Prop. of nodes in Giant Component

Figure 4.A6: Scaling of subnetwork statistics for empirical networks obtained by failing links. (a.) When all nodes are known q links are observed through sampling, the sample statistic for the number of nodes n equals the true number of nodes N. It should be noted, though, that some nodes of degree 0 may be observed and these are counted as nodes (not discarded). (b.) The number of edges scales linearly as $M_{\text{obs}} = qM$. (c.) The average degree scales linearly as $k_{\text{avg}}^{obs} = \frac{k_{\text{avg}}^{true}}{q}$. (d.) The max degree scales linearly (e.) Clustering scales roughly linearly with q. (f.) The percolation threshold roughly corresponds to the q for which $k_{\text{avg}} \geq 1$.



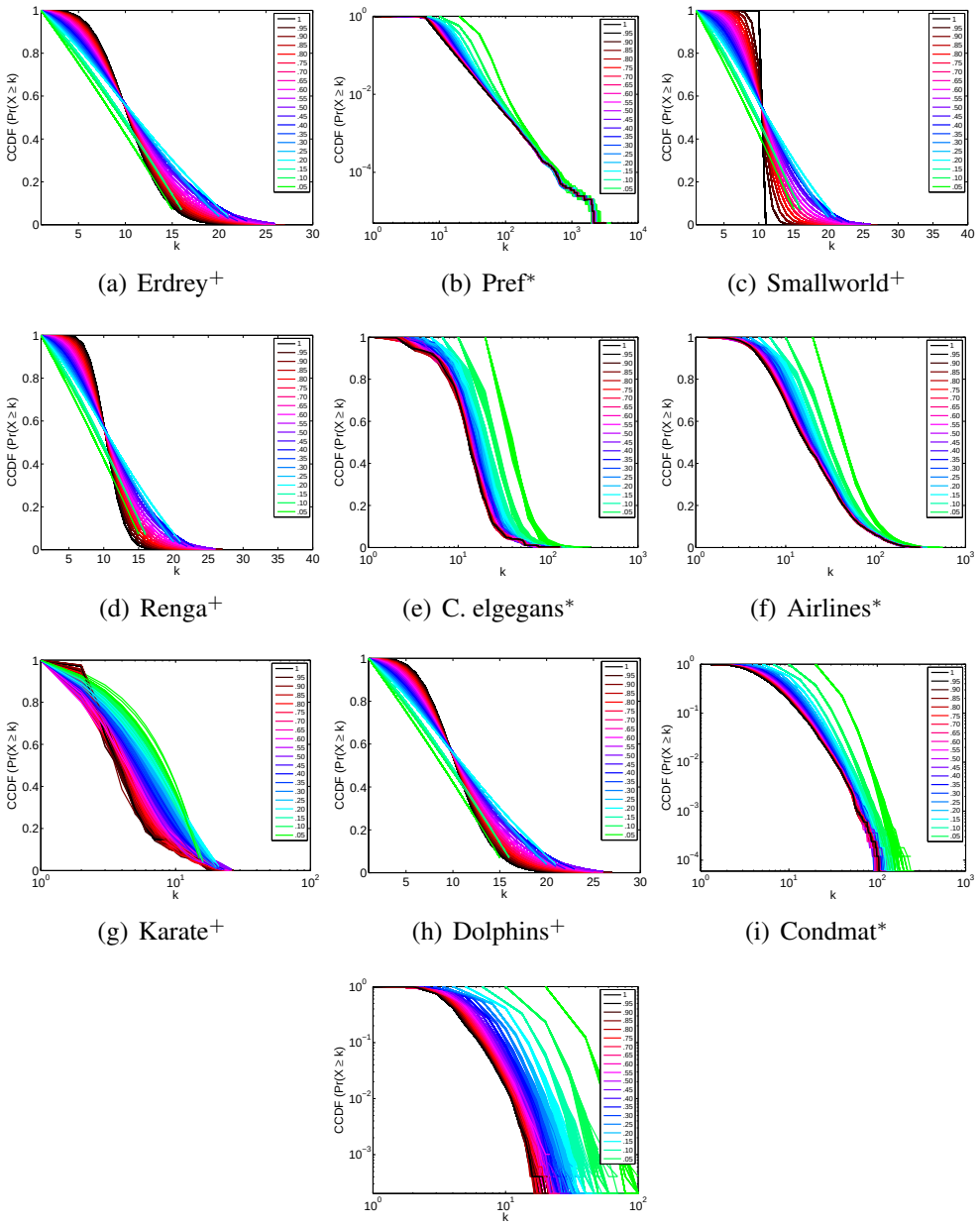
(a) Erdrey (b) Pref (c) Smallworld

(d) Renga (e) C. elgegans (f) Airlines

(g) Karate (h) Dolphins (i) Condmat

(j) Powergrid

Figure 4.A7: CCDF distortion for subnetworks obtained by failing links. Subnetwork degree distributions do not capture the true degree distribution, especially for small q.



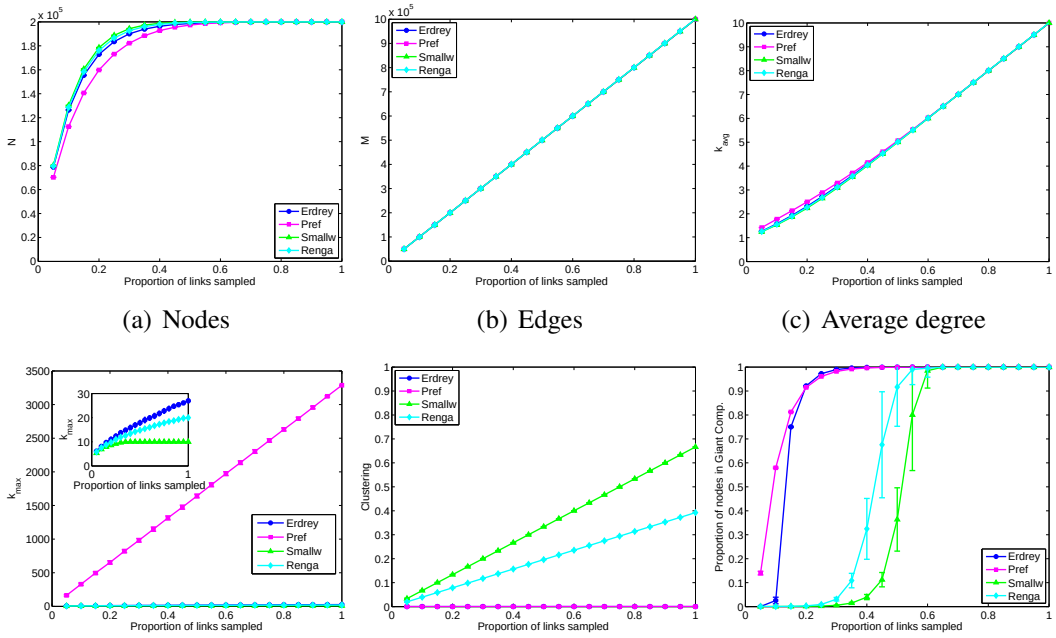
(a) Erdrey^{+} (b) Pref* (c) Smallworld^{+}

(d) Renga^{+} (e) C. elegans* (f) Airlines*

(g) Karate^{+} (h) Dolphins^{+} (i) Condmat*

(j) Powergrid*

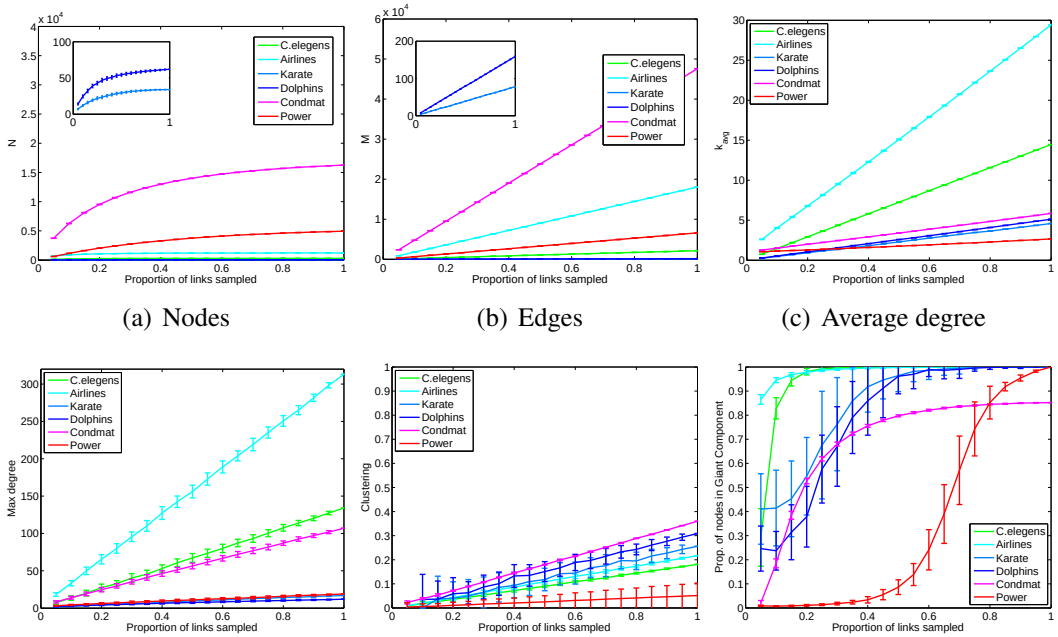
Figure 4.A8: Predicted CCDF from subnetworks obtained by failing links. The predicted CCDF shows relatively good agreement with the CCDF for most networks. Karate club and Dolphins exhibit significant deviations, possibly due to the small number of nodes in these networks. Networks designated with $^{+}$ utilized Equation 4.15 and those designated with with * utilized Equation 4.17.



(a) Nodes (b) Edges (c) Average degree

(d) Max degree (e) Clustering (f) Prop. of links in Giant Component

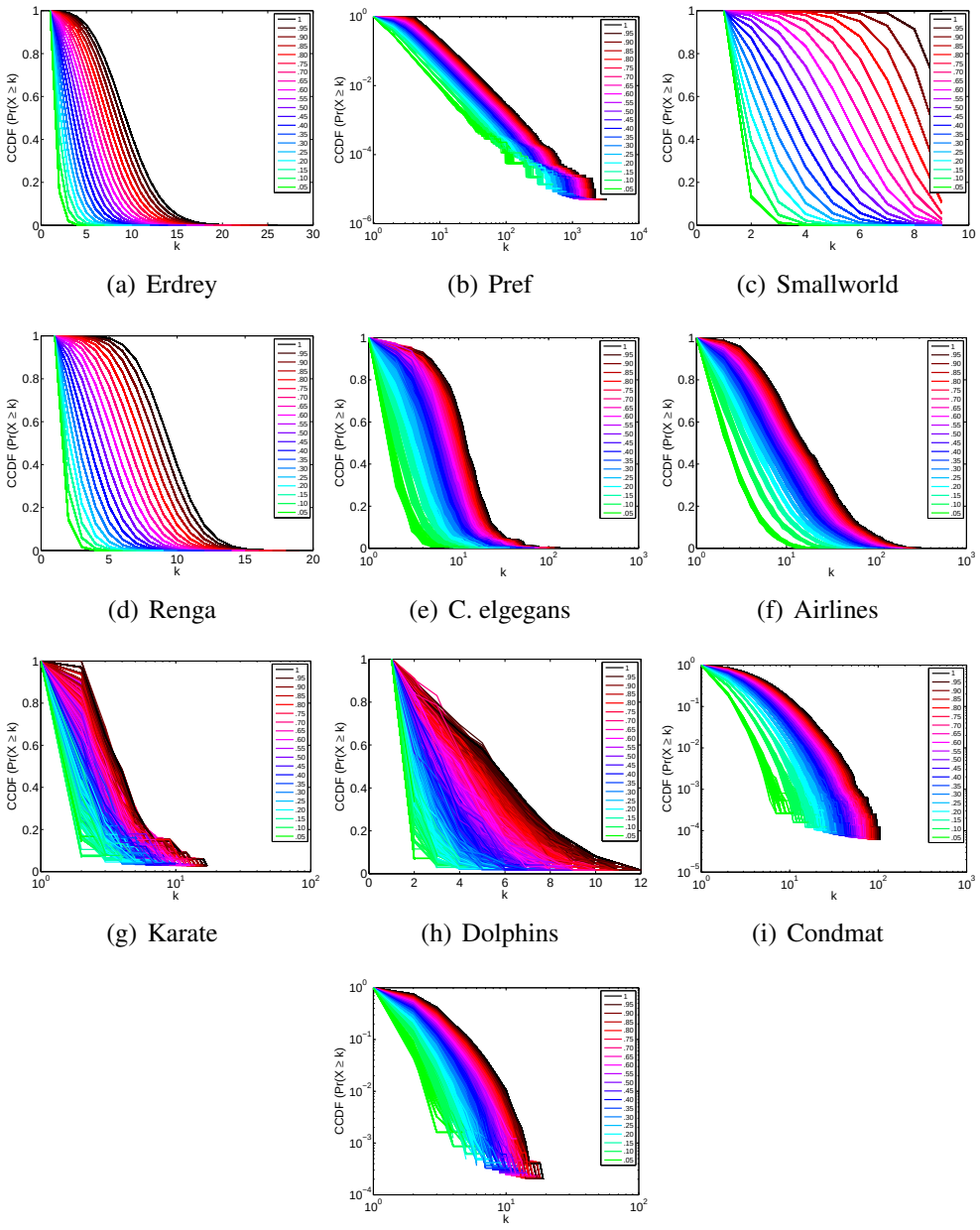
Figure 4.A9: Scaling of subnetwork statistics for simulated networks induced on sampled links. (a.) The number of nodes in a subnetwork sampled by links scales nonlinearly with q. (b.) The number of edges scales as $m \approx qM$. (c.) The average degree scales roughly linearly with the proportion of nodes subsampled $k_{\text{avg}}^{sub} \approx qk_{\text{avg}}$. (d.) The max degree scales roughly linearly for networks with few large hubs (e.g., Pref) and nonlinearly when there are several nodes with degrees roughly similar to $k_{\max}$. (e.) The clustering coefficient scales roughly linearly $C^{sub} \approx qC$. (f.) The proportion of nodes in the giant component increases with the proportion of nodes sampled. For the random graphs (Erdrey and Pref) there is a critical point corresponding to the approximate sampling level when $k_{\text{avg}} > 1$ (which corresponds to $q = 0.1$). The thresholds for Small World and Range dependent networks are much higher due to the uniformity of the motif distribution in these networks. Markers indicates the mean over 100 simulations. Error bars showing one standard deviation are too small to see.



(a) Nodes (b) Edges (c) Average degree

(d) Max degree (e) Clustering (f) Giant Component

Figure 4.A10: Scaling of subnetwork statistics for empirical networks induced on sampled links. (a.) The number of nodes in a subnetwork sampled by nodes scales nonlinearly with q. (b.) The number of edges scales as $m \approx qM$. (c.) The average degree scales roughly linearly with the proportion of nodes subsampled $k_{\text{avg}}^{sub} \approx qk_{\text{avg}}$. (d.) The max degree scales roughly linearly for networks with few large hubs. (e.) The clustering coefficient scales roughly linearly $C^{sub} \approx qC$. (f.) The proportion of nodes in the giant component increases with the proportion of links sampled. *C. elegans* and airlines maintain a large proportion of nodes in the giant component, most likely because these networks have high average degree. Karate club and dolphins show considerable variability (as shown by error bars $\pm$ s.d.) because these are relatively small networks. Powergrid is fragile to sampling by links, meaning the a high proportion of sampled links must be obtained to reach a fully connected network.



(a) Erdrey (b) Pref (c) Smallworld

(d) Renga (e) C. elgegans (f) Airlines

(g) Karate (h) Dolphins (i) Condmat

(j) Powergrid

Figure 4.A11: CCDF distortion for subnetworks induced on sampled links. Subnetwork degree distributions do not capture the true degree distribution, especially for small q.

Table 4.A1: Error in $\hat{N}$ when sampling by nodes.

q	Erdrey	Pref	Smallw	Renga	C.elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A2: Error in $\hat{M}$ when sampling by nodes. The percent error in the number of predicted nodes is nearly zero when, except in the small empirical networks where for small q, we violate the assumption that $n >> 1$ and incur large errors.

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.00	0.00	0.00	0.00	0.08	0.02	2.71	2.04	0.00	0.01
0.10	0.00	0.00	0.00	0.00	0.02	0.03	1.04	0.28	0.00	0.00
0.15	0.00	0.00	0.00	0.00	0.01	0.00	0.24	0.04	0.00	0.01
0.20	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.02	0.00	0.00
0.25	0.00	0.00	0.00	0.00	0.06	0.01	0.08	0.06	0.01	0.00
0.30	0.00	0.00	0.00	0.00	0.02	0.03	0.05	0.02	0.00	0.00
0.35	0.00	0.00	0.00	0.00	0.02	0.01	0.02	0.04	0.00	0.00
0.40	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00
0.45	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.03	0.00	0.00
0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.01	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.03	0.00	0.00
0.60	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.00	0.00
0.65	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00
0.70	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.02	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.01	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A3: Error in $\hat{k}_{\text{avg}}$ when sampling by nodes. Errors in $\hat{M}$ are largely responsible for errors in $\hat{k}_{\text{avg}}$.

q	Erdrey	Pref	Smallw	Renga	C.elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.00	0.00	0.00	0.00	0.08	0.02	2.71	2.04	0.00	0.01
0.10	0.00	0.00	0.00	0.00	0.02	0.03	1.04	0.28	0.00	0.00
0.15	0.00	0.00	0.00	0.00	0.01	0.00	0.24	0.04	0.00	0.01
0.20	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.02	0.00	0.00
0.25	0.00	0.00	0.00	0.00	0.06	0.01	0.08	0.06	0.01	0.00
0.30	0.00	0.00	0.00	0.00	0.02	0.03	0.05	0.02	0.00	0.00
0.35	0.00	0.00	0.00	0.00	0.02	0.01	0.02	0.04	0.00	0.00
0.40	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00
0.45	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.03	0.00	0.00
0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.01	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.03	0.00	0.00
0.60	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.00	0.00
0.65	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00
0.70	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.02	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.01	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A4: Error in $\hat{k}_{\max}$ when sampling by nodes. The percent error in the predicted max degree is nearly zero for large q. In general, predicting the max. degree is difficult due to the dependence on network structure.

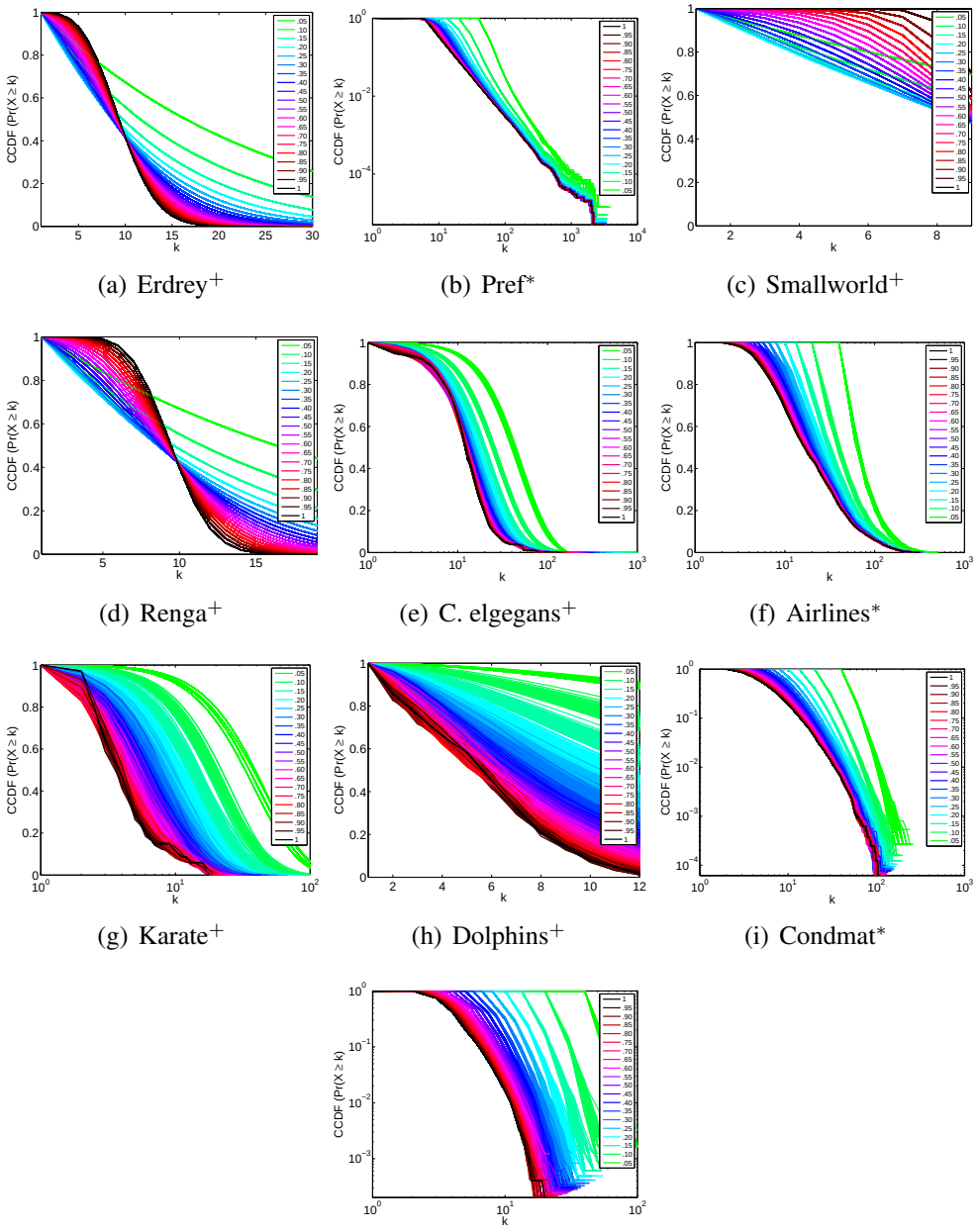
q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	2.70	0.67	7.70	3.73	0.59	0.39	0.00	0.08	0.13	0.14
0.10	1.60	0.54	4.94	2.26	0.52	0.28	0.18	0.02	0.09	0.08
0.15	1.13	0.49	3.73	1.67	0.46	0.21	0.31	0.01	0.05	0.05
0.20	0.89	0.42	2.96	1.29	0.46	0.18	0.30	0.01	0.06	0.04
0.25	0.72	0.38	2.46	1.06	0.44	0.15	0.28	0.01	0.05	0.03
0.30	0.57	0.33	2.09	0.87	0.33	0.12	0.21	0.01	0.05	0.02
0.35	0.48	0.27	1.77	0.73	0.33	0.12	0.18	0.01	0.06	0.01
0.40	0.40	0.24	1.50	0.62	0.30	0.09	0.10	0.01	0.05	0.01
0.45	0.34	0.21	1.22	0.52	0.25	0.07	0.17	0.01	0.02	0.01
0.50	0.29	0.19	1.00	0.44	0.20	0.07	0.13	0.01	0.01	0.01
0.55	0.24	0.16	0.82	0.38	0.20	0.07	0.11	0.01	0.01	0.01
0.60	0.21	0.15	0.67	0.33	0.19	0.04	0.04	0.01	0.00	0.01
0.65	0.17	0.13	0.54	0.27	0.16	0.04	0.03	0.01	0.00	0.00
0.70	0.14	0.10	0.43	0.23	0.10	0.04	0.02	0.00	0.01	0.00
0.75	0.11	0.07	0.33	0.18	0.12	0.04	0.01	0.00	0.01	0.00
0.80	0.09	0.05	0.25	0.13	0.10	0.02	0.01	0.00	0.01	0.00
0.85	0.07	0.04	0.18	0.10	0.07	0.01	0.01	0.00	0.00	0.00
0.90	0.04	0.03	0.11	0.07	0.04	0.01	0.00	0.00	0.00	0.00
0.95	0.02	0.02	0.05	0.04	0.03	0.01	0.01	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A5: Error in $\hat{C}$ when sampling by nodes. For some small networks with a small portion of nodes sampled q, no paths of length three occurred and the clustering coefficient was not computed in these cases.

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.01	0.33	0.00	0.00	–	0.04	–	–	0.00	–
0.10	0.07	0.15	0.00	0.00	0.09	0.02	–	–	0.01	0.21
0.15	0.04	0.10	0.00	0.00	0.08	0.02	–	–	0.01	0.03
0.20	0.03	0.07	0.00	0.00	0.01	0.02	–	–	0.01	0.02
0.25	0.05	0.06	0.00	0.00	0.01	0.01	–	–	0.00	0.07
0.30	0.04	0.05	0.00	0.00	0.05	0.01	–	0.15	0.00	0.03
0.35	0.05	0.03	0.00	0.00	0.03	0.01	–	0.14	0.00	0.06
0.40	0.05	0.02	0.00	0.00	0.00	0.01	0.06	0.12	0.00	0.02
0.45	0.03	0.02	0.00	0.00	0.01	0.01	0.13	0.02	0.00	0.05
0.50	0.01	0.02	0.00	0.00	0.01	0.01	0.20	0.04	0.00	0.02
0.55	0.00	0.01	0.00	0.00	0.01	0.00	0.12	0.05	0.00	0.00
0.60	0.02	0.00	0.00	0.00	0.01	0.00	0.08	0.02	0.00	0.02
0.65	0.01	0.00	0.00	0.00	0.01	0.00	0.04	0.01	0.00	0.01
0.70	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.01	0.00	0.00
0.75	0.00	0.01	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00
0.80	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.85	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00
0.90	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A6: Error in $\hat{N}$ when sampling by failing links. No error is encountered because all nodes remain in the subnetwork.

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00



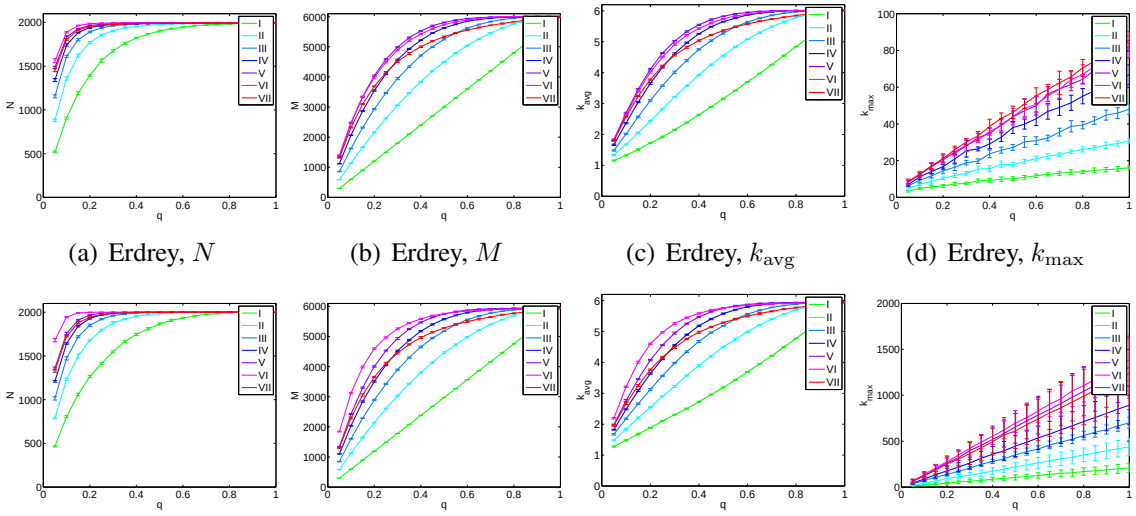
(a) Erdrey^{+} (b) Pref* (c) Smallworld^{+}

(d) Renga^{+} (e) C. elegans^{+} (f) Airlines*

(g) Karate^{+} (h) Dolphins^{+} (i) Condmat*

(j) Powergrid*

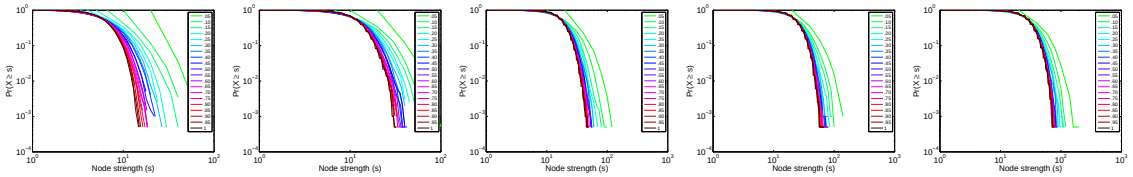
Figure 4.A12: Predicted CCDF from subnetworks induced on sampled links. The predicted CCDF shows relatively good agreement with the CCDF for most networks. Karate club and Dolphins exhibit significant deviations, possibly due to the small number of nodes in these networks. Networks designated with $^{+}$ utilized Equation 4.15 and those designated with with * utilized Equation 4.17.



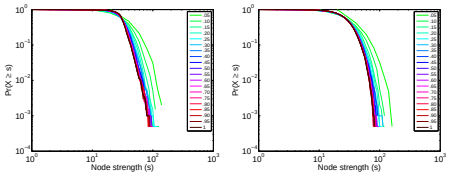
(a) Erdrey, N (b) Erdrey, M (c) Erdrey, k_{avg} (d) Erdrey, $k_{\max}$

(e) Pref, N (f) Pref, M (g) Pref, k_{avg} (h) Pref, $k_{\max}$

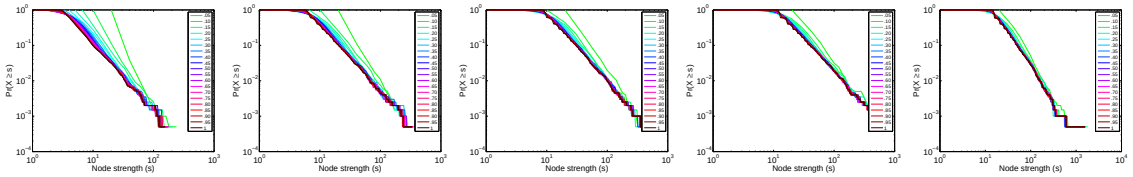
Figure 4.A13: Scaling of subnetwork statistics for simulated networks induced on sampled interactions.



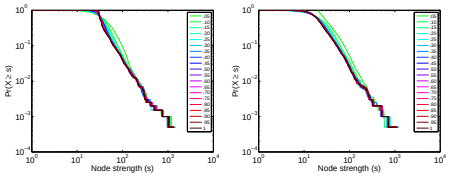
(a) Erdrey, Case 1 (b) Erdrey, Case 2 (c) Erdrey, Case 3 (d) Erdrey, Case 4 (e) Erdrey, Case 5



(f) Erdrey, Case 6 (g) Erdrey, Case 7

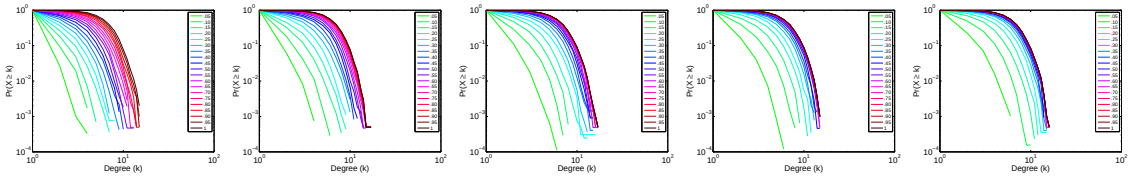


(h) Pref, Case 1 (i) Pref, Case 2 (j) Pref, Case 3 (k) Pref, Case 4 (l) Pref, Case 5

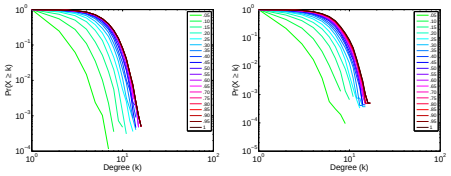


(m) Pref, Case 6 (n) Pref, Case 7

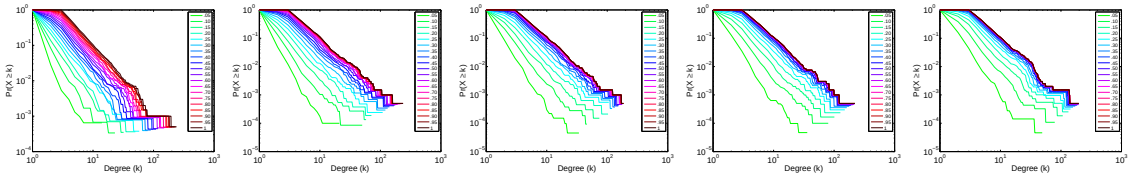
Figure 4.A14: Predicted node strength distribution for weighted, simulated networks.



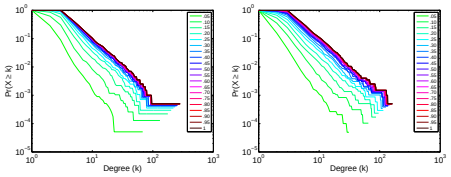
(a) Erdrey, Case 1 (b) Erdrey, Case 2 (c) Erdrey, Case 3 (d) Erdrey, Case 4 (e) Erdrey, Case 5



(f) Erdrey, Case 6 (g) Erdrey, Case 7

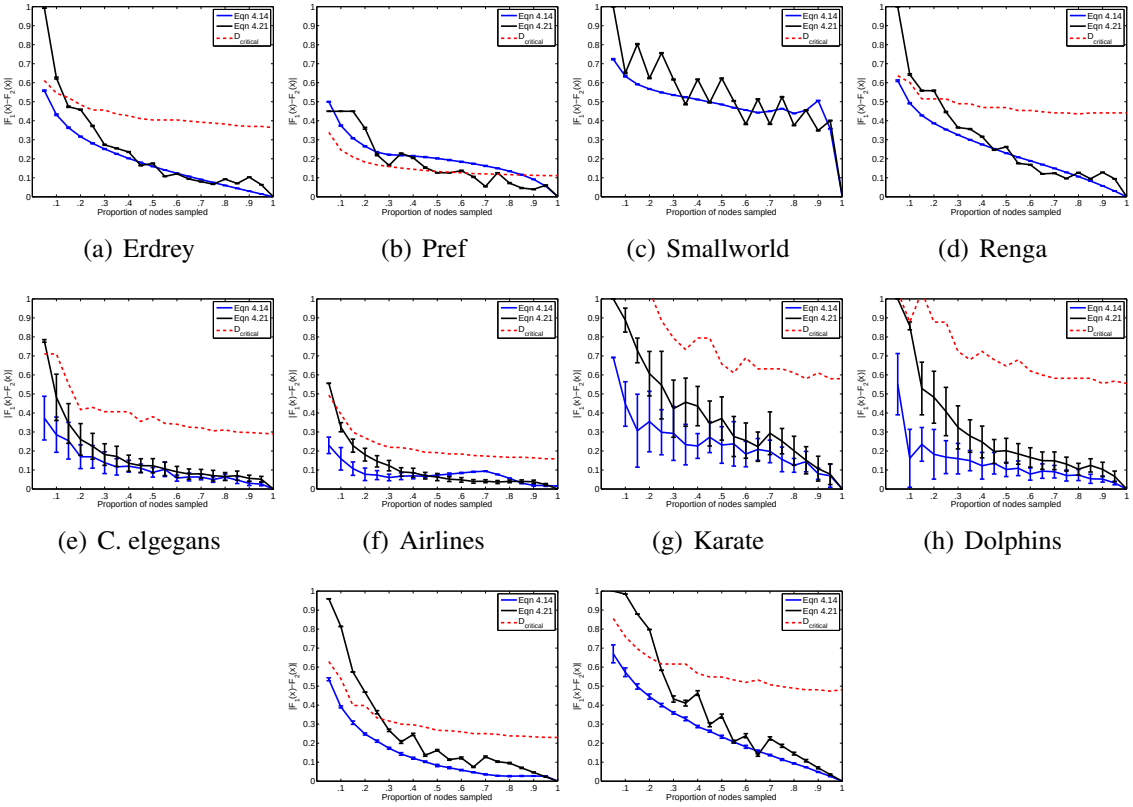


(h) Pref, Case 1 (i) Pref, Case 2 (j) Pref, Case 3 (k) Pref, Case 4 (l) Pref, Case 5



(m) Pref, Case 6 (n) Pref, Case 7

Figure 4.A15: Predicted degree distribution for weighted, simulated networks.



(a) Erdrey (b) Pref (c) Smallworld (d) Renga

(e) C. elegans (f) Airlines (g) Karate (h) Dolphins

(i) Condmat (j) Powergrid

Figure 4.A16: Kolmogorov-Smirnov two sample test for true CDF and predicted CDF from subnetworks induced on sampled nodes. The red line represents D_{crit} for $\alpha = 0.05$ and sample sizes $n_1 = k_{\max}$ of the true CDF and $n_2 = k_{\max}$ of the observed CDF. The predicted CDFs for for most networks are statistically indistinguishable from the true CDF for these networks for $q > 0.3$. Due to the presence of large hubs in Pref, n_1, n_2 are quite large leading to high statistical power in the KS test. Thus, even very small differences between the true and predicted CDFs result in a statistically significant difference and rejection of the null hypothesis, even though the curves show relatively good agreement.

Table 4.A7: Error in $\hat{M}$ when sampling by failing links. Since we are sampling qM links, errors in predicting the true number of links are quite small and nonzero only due to round-off error (e.g., m =round(qM)).

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00
0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A8: Error in $\hat{k}_{\text{avg}}$ when sampling by failing links. The predicted average degree is computed from $\hat{N}$ and $\hat{M}$. Error in the predicted average agree are small and only occur due to rounding errors in the selecting an integer number of qM edges in the random sample.

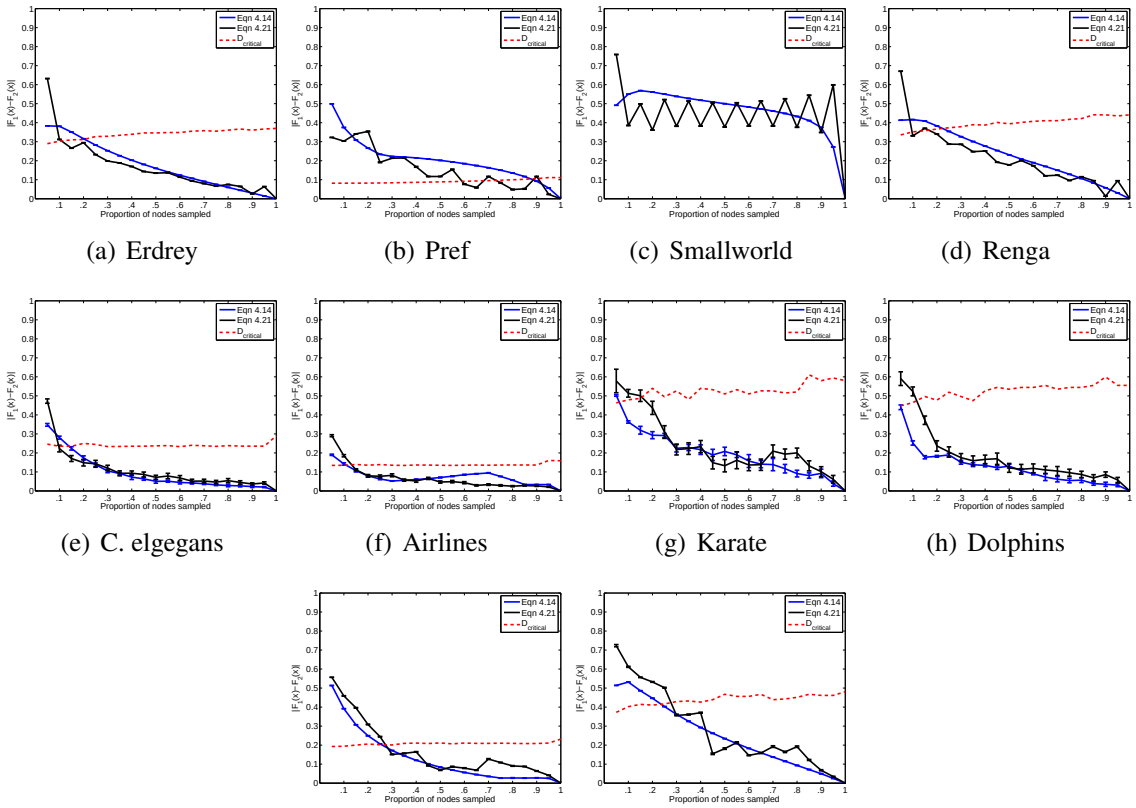
q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.00	0.00	0.00	0.00	0.03	0.03	0.06	0.09	0.00	0.00
0.10	0.00	0.00	0.00	0.00	0.02	0.02	0.03	0.05	0.00	0.00
0.15	0.00	0.00	0.00	0.00	0.02	0.01	0.03	0.03	0.00	0.00
0.20	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.02	0.00	0.00
0.25	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.02	0.00	0.00
0.30	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.00	0.00
0.35	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00
0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A9: Error in $\hat{C}$ when sampling by failing links.

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.50	0.11	0.00	0.01	0.13	0.18	–	–	0.06	0.28
0.10	0.66	0.05	0.00	0.00	0.05	0.03	0.36	0.04	0.02	0.18
0.15	0.18	0.02	0.00	0.00	0.00	0.01	0.18	0.24	0.06	0.05
0.20	0.06	0.02	0.00	0.00	0.00	0.04	0.45	0.18	0.03	0.01
0.25	0.05	0.00	0.00	0.00	0.02	0.00	0.01	0.09	0.10	0.04
0.30	0.03	0.01	0.00	0.00	0.01	0.00	0.21	0.02	0.03	0.00
0.35	0.01	0.01	0.00	0.00	0.01	0.01	0.02	0.02	0.01	0.02
0.40	0.02	0.01	0.00	0.00	0.01	0.01	0.15	0.07	0.00	0.01
0.45	0.01	0.01	0.00	0.00	0.01	0.01	0.05	0.02	0.04	0.00
0.50	0.01	0.00	0.00	0.00	0.00	0.01	0.07	0.01	0.02	0.01
0.55	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.01
0.60	0.00	0.00	0.00	0.00	0.01	0.00	0.05	0.03	0.00	0.00
0.65	0.00	0.01	0.00	0.00	0.00	0.01	0.06	0.02	0.00	0.00
0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.01	0.00
0.75	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.00
0.80	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00
0.95	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A10: Error in $\hat{k}_{\max}$ when sampling by failing links.

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.47	0.01	0.33	0.23	0.07	0.17	0.21	0.18	0.06	0.24
0.10	0.29	0.00	0.02	0.39	0.02	0.17	0.05	0.13	0.15	0.37
0.15	0.26	0.00	0.11	0.32	0.01	0.13	0.17	0.10	0.12	0.19
0.20	0.26	0.00	0.09	0.32	0.01	0.09	0.08	0.00	0.12	0.07
0.25	0.09	0.01	0.09	0.33	0.01	0.05	0.11	0.11	0.08	0.13
0.30	0.24	0.00	0.02	0.26	0.01	0.06	0.03	0.07	0.09	0.01
0.35	0.21	0.00	0.00	0.21	0.01	0.03	0.10	0.10	0.08	0.02
0.40	0.09	0.00	0.00	0.18	0.01	0.05	0.08	0.02	0.07	0.01
0.45	0.09	0.00	0.00	0.19	0.00	0.04	0.07	0.06	0.07	0.01
0.50	0.00	0.00	0.00	0.13	0.01	0.04	0.13	0.05	0.05	0.04
0.55	0.06	0.00	0.00	0.09	0.00	0.01	0.10	0.02	0.05	0.03
0.60	0.06	0.00	0.00	0.12	0.01	0.02	0.08	0.04	0.03	0.05
0.65	0.02	0.00	0.00	0.14	0.00	0.03	0.08	0.03	0.03	0.01
0.70	0.02	0.00	0.00	0.11	0.00	0.02	0.06	0.02	0.02	0.02
0.75	0.05	0.00	0.00	0.09	0.00	0.02	0.06	0.03	0.01	0.02
0.80	0.00	0.00	0.00	0.06	0.01	0.01	0.04	0.02	0.02	0.04
0.85	0.02	0.00	0.00	0.04	0.01	0.00	0.03	0.02	0.02	0.04
0.90	0.00	0.00	0.06	0.03	0.00	0.00	0.02	0.01	0.01	0.02
0.95	0.00	0.00	0.05	0.01	0.00	0.00	0.01	0.01	0.00	0.01
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00



(a) Erdrey (b) Pref (c) Smallworld (d) Renga

(e) C. elgegans (f) Airlines (g) Karate (h) Dolphins

(i) Condmat (j) Powergrid

Figure 4.A17: Kolmogorov-Smirnov two sample test for true CDF and predicted CDF from subnetworks obtained by failing links. The red line represents D_{crit} for $\alpha = 0.05$ and sample sizes $n_1 = k_{\max}$ of the true CDF and $n_2 = k_{\max}$ of the observed CDF. The predicted CDFs for for most networks are statistically indistinguishable from the true CDF for these networks for $q > 0.3$. Due to the presence of large hubs in Pref, n_1, n_2 are quite large leading to high statistical power in the KS test. Thus, even very small differences between the true and predicted CDFs result in a statistically significant difference and rejection of the null hypothesis, even though the curves show relatively good agreement.

Table 4.A11: Error in N when sampling by links. Predictors show good agreements with true values, except for low values of q. In these cases, errors in the predicted degree distribution contribute to errors in the predicted number of nodes. Future improvements in the predicted degree distribution would improve $\hat{N}$.

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.40	0.47	0.38	0.39	0.34	0.53	0.68	0.64	0.65	0.80
0.10	0.11	0.21	0.08	0.09	0.11	0.34	0.46	0.41	0.44	0.64
0.15	0.02	0.06	0.06	0.04	0.02	0.23	0.33	0.26	0.31	0.51
0.20	0.07	0.02	0.10	0.09	0.01	0.17	0.23	0.16	0.22	0.40
0.25	0.08	0.05	0.10	0.09	0.01	0.12	0.15	0.10	0.15	0.31
0.30	0.07	0.07	0.08	0.08	0.01	0.10	0.10	0.06	0.11	0.24
0.35	0.05	0.07	0.06	0.06	0.01	0.07	0.06	0.04	0.07	0.18
0.40	0.04	0.06	0.04	0.04	0.00	0.06	0.04	0.03	0.05	0.14
0.45	0.03	0.05	0.02	0.03	0.00	0.04	0.00	0.02	0.03	0.10
0.50	0.02	0.04	0.01	0.02	0.00	0.03	0.01	0.02	0.02	0.07
0.55	0.01	0.03	0.01	0.01	0.00	0.03	0.01	0.01	0.01	0.05
0.60	0.01	0.02	0.00	0.00	0.00	0.01	0.02	0.01	0.01	0.03
0.65	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.01	0.00	0.02
0.70	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01
0.75	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A12: Error in M when sampling by links. Error is nonzero only because of round-off errors when selecting an integer number of edges to sample.

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00
0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00
0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A13: Error in k_{avg} when sampling by links.

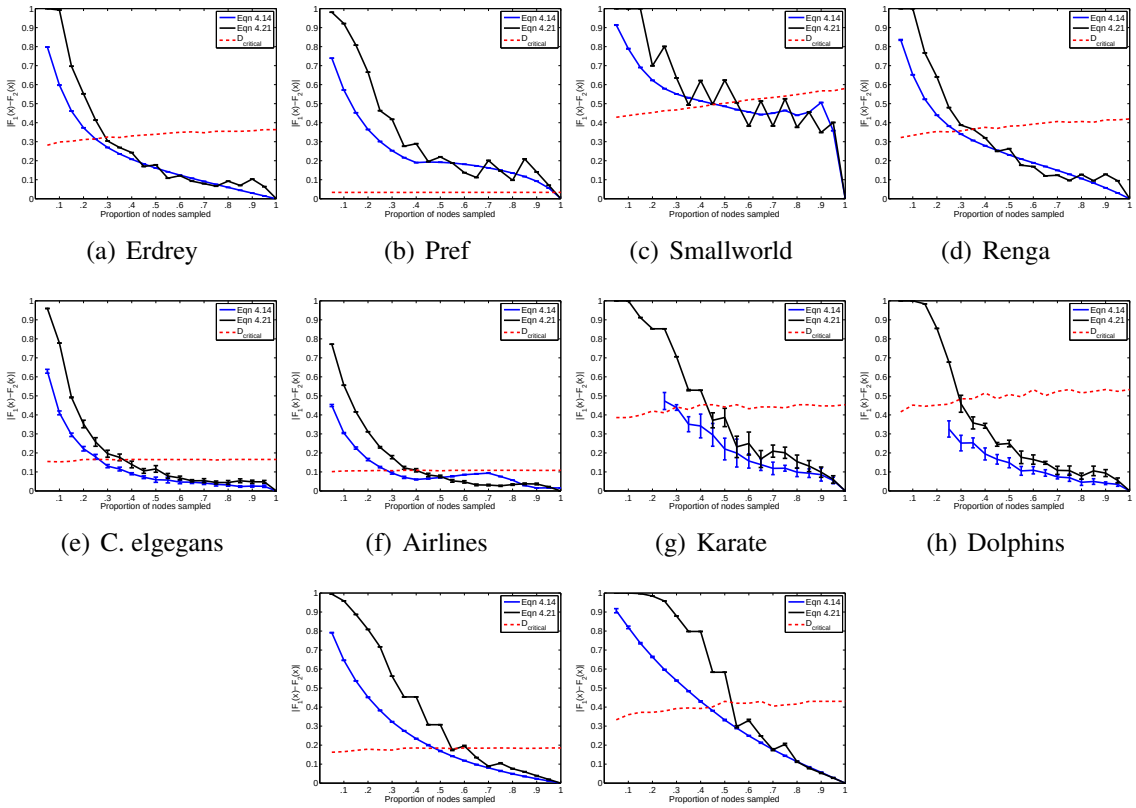
q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.66	0.89	0.61	0.63	0.50	1.13	2.18	1.79	1.83	4.03
0.10	0.12	0.26	0.08	0.10	0.12	0.51	0.90	0.71	0.79	1.79
0.15	0.02	0.07	0.05	0.04	0.02	0.30	0.52	0.36	0.45	1.04
0.20	0.06	0.02	0.09	0.08	0.01	0.20	0.32	0.20	0.28	0.67
0.25	0.07	0.05	0.09	0.08	0.01	0.14	0.21	0.12	0.18	0.46
0.30	0.06	0.06	0.07	0.07	0.01	0.11	0.09	0.07	0.12	0.32
0.35	0.05	0.06	0.05	0.05	0.01	0.08	0.06	0.05	0.08	0.22
0.40	0.04	0.05	0.04	0.04	0.00	0.06	0.04	0.03	0.05	0.16
0.45	0.03	0.04	0.02	0.03	0.00	0.05	0.00	0.03	0.03	0.11
0.50	0.02	0.03	0.01	0.02	0.00	0.03	0.01	0.02	0.02	0.07
0.55	0.01	0.02	0.01	0.01	0.00	0.03	0.01	0.00	0.01	0.05
0.60	0.01	0.02	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.03
0.65	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.02
0.70	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01
0.75	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A14: Error in C when sampling by links.

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.51	0.20	0.00	0.01	0.05	0.15	–	–	0.02	0.05
0.10	0.36	0.05	0.00	0.01	0.04	0.05	–	–	0.00	0.27
0.15	0.21	0.00	0.00	0.00	0.00	0.06	–	–	0.01	0.03
0.20	0.20	0.02	0.00	0.00	0.02	0.01	–	–	0.00	0.02
0.25	0.01	0.00	0.00	0.00	0.01	0.00	–	0.19	0.00	0.04
0.30	0.00	0.00	0.00	0.00	0.02	0.02	0.16	0.06	0.00	0.01
0.35	0.05	0.00	0.00	0.00	0.00	0.01	0.05	0.07	0.00	0.00
0.40	0.03	0.01	0.00	0.00	0.00	0.01	0.08	0.07	0.00	0.01
0.45	0.02	0.01	0.00	0.00	0.00	0.01	0.05	0.03	0.00	0.02
0.50	0.01	0.00	0.00	0.00	0.00	0.00	0.08	0.02	0.00	0.03
0.55	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.02	0.00	0.01
0.60	0.01	0.00	0.00	0.00	0.01	0.01	0.06	0.02	0.00	0.01
0.65	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
0.70	0.00	0.01	0.00	0.00	0.01	0.00	0.02	0.01	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.01	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.00	0.00
0.85	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A15: Error in $k_{\max}$ when sampling by links.

q	Erdrey	Pref	Smallw	Renga	C. elegans	Airlines	Karate	Dolphins	Condmat	Power
0.05	0.67	0.00	0.20	0.16	0.11	0.18	1.14	2.38	0.06	0.24
0.10	0.33	0.00	0.05	0.37	0.01	0.09	0.62	1.39	0.15	0.37
0.15	0.30	0.00	0.10	0.18	0.00	0.14	0.42	0.02	0.12	0.19
0.20	0.28	0.01	0.10	0.40	0.02	0.10	0.36	0.02	0.12	0.07
0.25	0.17	0.00	0.05	0.23	0.03	0.06	0.32	0.10	0.08	0.13
0.30	0.17	0.00	0.03	0.24	0.01	0.07	0.16	0.11	0.09	0.01
0.35	0.15	0.00	0.00	0.27	0.00	0.04	0.15	0.03	0.08	0.02
0.40	0.19	0.00	0.00	0.20	0.01	0.04	0.11	0.05	0.07	0.01
0.45	0.11	0.00	0.00	0.11	0.00	0.05	0.13	0.04	0.07	0.01
0.50	0.07	0.00	0.00	0.16	0.01	0.03	0.13	0.04	0.05	0.04
0.55	0.01	0.00	0.00	0.15	0.00	0.03	0.09	0.06	0.05	0.03
0.60	0.09	0.00	0.00	0.17	0.01	0.03	0.06	0.02	0.03	0.05
0.65	0.08	0.00	0.00	0.13	0.01	0.01	0.09	0.04	0.03	0.01
0.70	0.07	0.00	0.00	0.10	0.00	0.02	0.06	0.02	0.02	0.02
0.75	0.02	0.00	0.00	0.07	0.01	0.02	0.06	0.02	0.01	0.02
0.80	0.01	0.00	0.00	0.03	0.00	0.00	0.04	0.03	0.02	0.04
0.85	0.00	0.00	0.00	0.04	0.00	0.00	0.03	0.03	0.02	0.04
0.90	0.01	0.00	0.05	0.02	0.01	0.00	0.01	0.01	0.01	0.02
0.95	0.01	0.00	0.05	0.00	0.00	0.00	0.00	0.02	0.00	0.01
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00



(a) Erdrey (b) Pref (c) Smallworld (d) Renga

(e) C. elgegans (f) Airlines (g) Karate (h) Dolphins

(i) Condmat (j) Powergrid

Figure 4.A18: Kolmogorov-Smirnov two sample test for true CDF and predicted CDF from subnetworks generated by sampled links. The red line represents D_{crit} for $\alpha = 0.05$ and sample sizes $n_1 = k_{\max}$ of the true CDF and $n_2 = k_{\max}$ of the observed CDF. The predicted CDFs for for most networks are statistically indistinguishable from the true CDF for these networks for $q > 0.3$. Due to the presence of large hubs in Pref, n_1, n_2 are quite large leading to high statistical power in the KS test. Thus, even very small differences between the true and predicted CDFs result in a statistically significant difference and rejection of the null hypothesis, even though the curves show relatively good agreement.

Table 4.A16: Error in $\hat{N}$ when sampling by interactions on an Erdös-Rényi random graph.

q	I	II	III	IV	V	VI	VII
0.05	0.54	0.50	0.46	0.41	0.36	0.34	0.36
0.10	0.48	0.39	0.30	0.23	0.18	0.14	0.18
0.15	0.42	0.28	0.19	0.12	0.09	0.06	0.10
0.20	0.35	0.20	0.12	0.07	0.05	0.03	0.05
0.25	0.29	0.14	0.07	0.04	0.02	0.01	0.03
0.30	0.24	0.10	0.05	0.02	0.01	0.01	0.02
0.35	0.19	0.07	0.03	0.02	0.01	0.00	0.01
0.40	0.14	0.05	0.02	0.01	0.01	0.00	0.01
0.45	0.11	0.03	0.01	0.01	0.00	0.00	0.01
0.50	0.08	0.02	0.01	0.00	0.00	0.00	0.00
0.55	0.06	0.01	0.01	0.00	0.00	0.00	0.00
0.60	0.05	0.01	0.00	0.00	0.00	0.00	0.00
0.65	0.03	0.01	0.00	0.00	0.00	0.00	0.00
0.70	0.02	0.00	0.00	0.00	0.00	0.00	0.00
0.75	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.80	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.85	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A17: Error in $\hat{N}$ when sampling by interactions from a Scale-free weighted network.

q	I	II	III	IV	V	VI	VII
0.05	0.36	0.36	0.36	0.36	0.36	0.36	0.36
0.10	0.18	0.18	0.18	0.18	0.18	0.18	0.18
0.15	0.10	0.10	0.10	0.10	0.10	0.10	0.10
0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05
0.25	0.03	0.03	0.03	0.03	0.03	0.03	0.03
0.30	0.02	0.02	0.02	0.02	0.02	0.02	0.02
0.35	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.40	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.45	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A18: Error in $\hat{M}$ when sampling by interactions from an Erdös-Rényi weighted network.

q	I	II	III	IV	V	VI	VII
0.05	0.00	0.85	0.78	0.72	0.66	0.65	0.67
0.10	0.00	0.71	0.60	0.49	0.40	0.41	0.44
0.15	0.00	0.59	0.44	0.32	0.22	0.24	0.29
0.20	0.00	0.48	0.31	0.19	0.10	0.12	0.19
0.25	0.00	0.38	0.21	0.10	0.02	0.05	0.13
0.30	0.00	0.30	0.13	0.03	0.02	0.00	0.08
0.35	0.00	0.22	0.07	0.01	0.05	0.02	0.06
0.40	0.00	0.16	0.02	0.04	0.06	0.04	0.04
0.45	0.00	0.11	0.01	0.05	0.06	0.04	0.03
0.50	0.00	0.07	0.03	0.05	0.05	0.04	0.02
0.55	0.00	0.03	0.04	0.04	0.04	0.03	0.01
0.60	0.00	0.01	0.04	0.04	0.03	0.02	0.01
0.65	0.00	0.01	0.04	0.03	0.02	0.02	0.01
0.70	0.00	0.02	0.03	0.02	0.01	0.01	0.01
0.75	0.00	0.02	0.02	0.01	0.01	0.01	0.01
0.80	0.00	0.02	0.02	0.01	0.00	0.00	0.00
0.85	0.00	0.02	0.01	0.00	0.00	0.00	0.00
0.90	0.00	0.01	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A19: Error in $\hat{M}$ when sampling by interactions Scale-free weighted network.

q	I	II	III	IV	V	VI	VII
0.05	0.67	0.67	0.67	0.67	0.67	0.67	0.67
0.10	0.44	0.44	0.44	0.44	0.44	0.44	0.44
0.15	0.29	0.29	0.29	0.29	0.29	0.29	0.29
0.20	0.19	0.19	0.19	0.19	0.19	0.19	0.19
0.25	0.13	0.13	0.13	0.13	0.13	0.13	0.13
0.30	0.08	0.08	0.08	0.08	0.08	0.08	0.08
0.35	0.06	0.06	0.06	0.06	0.06	0.06	0.06
0.40	0.04	0.04	0.04	0.04	0.04	0.04	0.04
0.45	0.03	0.03	0.03	0.03	0.03	0.03	0.03
0.50	0.02	0.02	0.02	0.02	0.02	0.02	0.02
0.55	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.60	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.65	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.70	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.75	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A20: Error in $\hat{k}_{\text{avg}}$ when sampling by interactions from an Erdös-Rényi weighted network.

q	I	II	III	IV	V	VI	VII
0.05	0.35	0.90	0.85	0.80	0.75	0.74	0.76
0.10	0.33	0.79	0.69	0.59	0.49	0.48	0.53
0.15	0.29	0.68	0.53	0.40	0.28	0.28	0.35
0.20	0.26	0.57	0.38	0.24	0.14	0.14	0.23
0.25	0.23	0.46	0.26	0.13	0.04	0.06	0.15
0.30	0.19	0.36	0.17	0.05	0.01	0.01	0.10
0.35	0.16	0.27	0.10	0.00	0.04	0.02	0.07
0.40	0.13	0.20	0.04	0.03	0.05	0.03	0.05
0.45	0.10	0.14	0.01	0.04	0.05	0.04	0.03
0.50	0.08	0.09	0.02	0.05	0.04	0.04	0.02
0.55	0.06	0.05	0.03	0.04	0.03	0.03	0.02
0.60	0.04	0.02	0.04	0.04	0.03	0.02	0.01
0.65	0.03	0.00	0.03	0.03	0.02	0.02	0.01
0.70	0.02	0.01	0.03	0.02	0.01	0.01	0.01
0.75	0.01	0.02	0.02	0.01	0.00	0.01	0.01
0.80	0.01	0.02	0.01	0.01	0.00	0.00	0.00
0.85	0.01	0.01	0.01	0.00	0.00	0.00	0.00
0.90	0.00	0.01	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A21: Error in $\hat{k}_{\text{avg}}$ when sampling by interactions from a Scale-free weighted network.

q	I	II	III	IV	V	VI	VII
0.05	0.76	0.76	0.76	0.76	0.76	0.76	0.76
0.10	0.53	0.53	0.53	0.53	0.53	0.53	0.53
0.15	0.35	0.35	0.35	0.35	0.35	0.35	0.35
0.20	0.23	0.23	0.23	0.23	0.23	0.23	0.23
0.25	0.15	0.15	0.15	0.15	0.15	0.15	0.15
0.30	0.10	0.10	0.10	0.10	0.10	0.10	0.10
0.35	0.07	0.07	0.07	0.07	0.07	0.07	0.07
0.40	0.05	0.05	0.05	0.05	0.05	0.05	0.05
0.45	0.03	0.03	0.03	0.03	0.03	0.03	0.03
0.50	0.02	0.02	0.02	0.02	0.02	0.02	0.02
0.55	0.02	0.02	0.02	0.02	0.02	0.02	0.02
0.60	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.65	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.70	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.75	0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.A22: Error in $k_{\max}$ when sampling by interactions from an Erdös-Rényi weighted network.

q	I	II	III	IV	V	VI	VII
0.05	3.00	0.76	0.81	0.84	0.83	0.84	0.85
0.10	1.82	0.66	0.73	0.76	0.76	0.77	0.80
0.15	1.27	0.60	0.67	0.71	0.69	0.70	0.73
0.20	0.82	0.53	0.60	0.66	0.62	0.66	0.69
0.25	0.72	0.48	0.56	0.59	0.59	0.60	0.63
0.30	0.49	0.44	0.51	0.52	0.54	0.55	0.58
0.35	0.51	0.35	0.50	0.52	0.49	0.53	0.55
0.40	0.35	0.36	0.42	0.49	0.47	0.47	0.49
0.45	0.29	0.28	0.39	0.44	0.41	0.44	0.45
0.50	0.20	0.31	0.37	0.37	0.37	0.39	0.42
0.55	0.17	0.26	0.32	0.36	0.34	0.34	0.37
0.60	0.16	0.22	0.33	0.33	0.31	0.32	0.33
0.65	0.13	0.20	0.31	0.28	0.23	0.27	0.30
0.70	0.10	0.19	0.26	0.26	0.20	0.23	0.26
0.75	0.08	0.15	0.21	0.23	0.17	0.17	0.23
0.80	0.01	0.13	0.21	0.18	0.14	0.14	0.18
0.85	0.02	0.12	0.17	0.14	0.08	0.08	0.14
0.90	0.01	0.09	0.12	0.11	0.04	0.04	0.11
0.95	0.04	0.08	0.10	0.06	0.01	0.02	0.06
1.00	0.05	0.04	0.06	0.02	0.07	0.05	0.03

Table 4.A23: Error in $k_{\max}$ when sampling by interactions from a Scale-free weighted network.

q	I	II	III	IV	V	VI	VII
0.05	0.85	0.85	0.85	0.85	0.85	0.85	0.85
0.10	0.80	0.80	0.80	0.80	0.80	0.80	0.80
0.15	0.73	0.73	0.73	0.73	0.73	0.73	0.73
0.20	0.69	0.69	0.69	0.69	0.69	0.69	0.69
0.25	0.63	0.63	0.63	0.63	0.63	0.63	0.63
0.30	0.58	0.58	0.58	0.58	0.58	0.58	0.58
0.35	0.55	0.55	0.55	0.55	0.55	0.55	0.55
0.40	0.49	0.49	0.49	0.49	0.49	0.49	0.49
0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
0.50	0.42	0.42	0.42	0.42	0.42	0.42	0.42
0.55	0.37	0.37	0.37	0.37	0.37	0.37	0.37
0.60	0.33	0.33	0.33	0.33	0.33	0.33	0.33
0.65	0.30	0.30	0.30	0.30	0.30	0.30	0.30
0.70	0.26	0.26	0.26	0.26	0.26	0.26	0.26
0.75	0.23	0.23	0.23	0.23	0.23	0.23	0.23
0.80	0.18	0.18	0.18	0.18	0.18	0.18	0.18
0.85	0.14	0.14	0.14	0.14	0.14	0.14	0.14
0.90	0.11	0.11	0.11	0.11	0.11	0.11	0.11
0.95	0.06	0.06	0.06	0.06	0.06	0.06	0.06
1.00	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Table 4.A24: Number of messages from September 2008-November 2009. The number of "observed" messages in our database comprise a fraction of the total number of Twitter messages made during period of this study (September 2008 through November 2009). While our feed from the Twitter API remains fairly constant, the total # of tweets grows, thus reducing the % of all tweets observed in our database. We calculate the total # of messages as the difference between the last message id and the first message id that we observe for a given month. This provides a reasonable estimation of the number of tweets made per month as message ids were assigned (by Twitter) sequentially during the time period of this study. The % observed represent the percent of messages observed out of the estimated total. We also report the number observed messages that are replies to specific messages and the percentage of our observed messages which constitute replies.

Week	Start date	# Obsvd. Msgs. $\times 10^6$	# Total Msgs. $\times 10^6$	% Obsvd.	# Replies $\times 10^6$	% Replies
1	09.09.08	3.14	7.26	43.2	0.88	28.1
2	09.16.08	3.36	8.31	40.4	0.90	26.9
3	09.23.08	3.43	8.89	38.6	0.90	26.2
4	09.30.08	3.33	9.06	36.8	0.89	26.6
5	10.07.08	2.33	9.38	24.8	0.64	27.5
6	10.14.08	4.39	9.87	44.4	1.24	28.3
7	10.21.08	4.70	10.01	47.0	1.35	28.8
8	10.28.08	5.74	10.34	55.5	1.64	28.5
9	11.04.08	5.58	11.14	50.1	1.63	29.3
10	11.11.08	4.70	9.88	47.6	1.42	30.2

Chapter 5

Conclusion

In this work, we describe the construction of Twitter reply and reciprocal reply networks. Countering claims that Twitter is not social a network (Kwak, et al., 2010), we provide evidence of a social subnetwork structure within Twitter. Given that our networks are derived from only a fraction of all tweets authored during the weeks under analysis, we are motivated to develop scaling methods to more accurately portray the global network statistics characterizing these networks. This analysis leads us to consider previous work, which is largely based on subnetworks induced on randomly selected nodes.

Subnetworks generated from randomly selected links differ substantially from those generated by randomly selected nodes. Most notably, the nodes in the former are a biased subsample in that hubs are much more likely to be included in the subnetwork. Because of this bias, Horvitz-Thompson estimators are required for predicting the number of nodes in the true network and this requires knowledge or the true degree distribution. Previous work has been challenged by this requirement. We surmount this challenge by providing a practical means of approximating the degree distribution. Using this approximation, we show that the Horvitz-Thompson estimators perform reasonably well for the true network statistics. We extend these methods to account for weighted networks and weighted, directed networks - sampling strategies largely unexplored in the literature. We conclude by providing estimates of the global network statistics for Twitter reply networks during the weeks from September 2008-November 2009.

The large volume of replies (millions every week) and assortativity of user happiness indicates that Twitter is being used as a social service. Furthermore, we find evidence of an upper threshold of approximately 150 neighbors. This supports previous work by Dunbar (1992), who found a positive correlation between the size of the neocortex of nonhuman primates and the number of social relationships that they can maintain. His theory suggests that humans can maintain approximately 150 social relationships. More recent support for Dunbar's number was detected in the work of Gonçalves et al. (2011). These researchers examine Twitter reply networks constructing from the Twitter firehose and find that edge weights of out-going edges gradually increase to a maximum around $k_{out} \approx 150$. They suggest that the "economy of attention" is a limiting factor restricting increased edge weights beyond this value.

Previous network constructions of Twitter utilized follower networks and our work overcomes the limitations of stale accumulation of social ties with no functional activity by examining an "in the moment" social network. By examining networks constructed at the time scale of weeks, we are able to view Twitter reciprocal reply networks as a dynamic social network. In this light, we examine a fundamental property of dynamic social networks: network densification. Using an evolutionary algorithm that exhibits fast convergence for optimizing real valued functions, we explore a link predictor that performs relatively well to other efforts in this realm. We note that a limitation of our work is the assumption of a linear model, as well as our inclusion of a highly unbalanced class for training our predictor. Future work which utilizes balanced classes, while still capitalizing on sparse matrix computations may be particularly fruitful. At the very least, such a development would enable a fair comparison of our method with state of the art supervised

learning methods, such as binary decision trees (with balanced classes). Additional work may explore the persistence or decay of links over time.

One of the most intriguing aspects of this work is the detection of similarity indices which evolve to have large, positive weights in our link predictors. Perhaps the most notable similarity index for which this is the case is the Resource Allocation Index. Resource allocation considers the amount of resource one node has and assumes that each node will distribute its resource equally among all neighbors (Zhou, Lü, & Zhang, 2009). Considering the limits to time and attention an individual has, this may be suggestive of a mechanism by which users limit their interaction.

While this work does not attempt to separate homophily and contagion, future work could examine the change in individuals' hedonometric scores, relative to changes in their nearest neighbors' and ambient hedonometric scores. Further research may explore the extent to which information or expressed sentiment flows in emergent virtual communities. Granovetter (1973) explores the role of strong and weak ties mediating the flow of information. Recent work by Weng, Menczer, and Ahn (2013) suggests that highly connected communities may trap contagion, however, the role of network topology has not been fully explored and it is also possible that communication of similar interests drives the evolution of tightly bound communities.

In a larger context, this work not only reveals a social network structure of Twitter, but also presents several tools for working with large, possibly incomplete network datasets. Future work which continues to explore techniques for inferring network topology from noisy or incomplete observations, as well as work which explore the role of topology on complex contagion will greatly advance our understanding behavior of networked systems.

References

Adamic, L. A. and E. Adar (2003). Friends and neighbors on the web. *Social Networks 25*(3), 211 – 230.

Adamic, L. A. and N. Glance (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link discovery*, pp. 36–43. ACM.

Aiello, L. M., A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web 6*(2), 9:1–9:33.

Al Hasan, M., V. Chaoji, S. Salem, and M. Zaki (2006). Link prediction using supervised learning. In *SDM06: Workshop on Link Analysis, Counter-terrorism and Security*.

Auger, A. and N. Hansen (2005). A restart CMA evolution strategy with increasing population size. In *IEEE Congress on Evolutionary Computation*, Volume 2, pp. 1769–1776. IEEE.

Backstrom, L. and J. Leskovec (2011). Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pp. 635–644. ACM.

Bagrow, J. P., S. Desu, M. R. Frank, N. Manukyan, L. Mitchell, A. Reagan, E. E. Bloedorn, L. B. Booker, L. K. Branting, M. J. Smith, B. F. Tivnan, C. M. Danforth, P. S. Dodds, and J. C. Bongard (2013). Shadow networks: Discovering hidden nodes with models of information flow. *arXiv preprint, arXiv:1312.6122*.

Bakshy, E., J. M. Hofman, W. A. Mason, and D. J. Watts (2011). Everone's an influencer: Quantifying influence on Twitter. In *WSDM '11: Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, New York, NY, USA.

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

Barabási, A.-L., H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications 311*(3), 590–614.

Barrat, A., M. Barthelemy, R. Pastor-Satorras, and A. Vespignani (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America 101*(11), 3747–3752.

Barrat, A., M. Barthlemy, and A. Vespignani (2008). *Dynamical processes on complex networks*. Cambridge University Press.

Bastian, M., H. Sebastien, and J. Mathieu (2009). Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*.

Biernacki, P. and D. Waldorf (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods and Research 10*(2), 141–163.

Bliss, C. A., M. R. Frank, C. M. Danforth, and P. S. Dodds (2014). An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science*.

Bliss, C. A., I. M. Kloumann, K. D. Harris, C. M. Danforth, and P. S. Dodds (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science 3*(5), 388 – 397.

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

Bollen, J., B. Goncalves, G. Ruan, and H. Mao (2011). Happiness is assortative in online social networks. *Artificial Life 17*(3).

Bollen, J., H. Mao, and X. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science 2*(1), 1–8.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2*(2), 121–167.

Cacioppo, J. T., J. H. Fowler, and N. A. Christakis (2009). Alone in the crowd: The structure and spread of loneliness in a large social network. *Journal of Personality and Social Psychology 97*(6), 977.

Cha, M., H. Haddadi, F. Benevenuto, and P. K. Gummadi (2010). Measuring user influence in Twitter: The million follower fallacy.

Chawla, N. V., N. Japkowicz, and A. Kotcz (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter 6*(1), 1–6.

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

Christakis, N. A. and J. H. Fowler (2008). The collective dynamics of smoking in a large social network. *New England Journal of Medicine 358*(21), 2249–2258.

Christakis, N. A. and J. H. Fowler (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine 32*, 556–577.

Cieslak, D. A. and N. V. Chawla (2008). Learning decision trees for unbalanced data. In *Machine Learning and Knowledge Discovery in Databases*, pp. 241–256. Springer.

Clauset, A., C. Shalizi, and M. Newman (2009). Power-law distributions in empirical data. *SIAM Review 51*(4), 661–703.

Cohen, R., K. Erez, D. Ben-Avraham, and S. Havlin (2000). Resilience of the internet to random breakdowns. *Physical Review Letters 85*(21), 4626.

Costenbader, E. and T. W. Valente (2003). The stability of centrality measures when networks are sampled. *Social Networks 25*(4), 283–307.

de Silva, E., T. Thorne, P. Ingram, I. Agrafioti, J. Swire, C. Wiuf, and M. Stumpf (2006). The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biology 4*(1), 39.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science 149*(3683), 510–515.

Dodds, P. S. and C. M. Danforth (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies 11*, 441–456.

Dodds, P. S., K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS one 6*(12), e26752.

Dunbar, R. I. M. (1995). Neocortex size and group size in primates: A test of the hypothesis. *Journal of Human Evolution 28*(3), 287 – 296.

Erdös, P. and A. Rényi (1960). On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl 5*, 17–61.

Esslimani, I., A. Brun, and A. Boyer (2011). Densifying a behavioral recommender system by social networks link prediction methods. *Social Network Analysis and Mining 1*(3), 159–172.

Fischer, E. and A. R. Reuber (2011). Social interaction via new social media: How can interactions on Twitter affect effectual thinking and behavior? *Journal of Business Venturing 26*(1), 1–18.

Fowler, J. H. and N. A. Christakis (2008). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ 337*.

Frank, M. R., L. Mitchell, P. S. Dodds, and C. M. Danforth (2013). Happiness and the patterns of life: A study of geolocated tweets. *Nature Scientific Reports 3*.

Frank, O. (19781979). Sampling and estimation in large social networks. *Social Networks 1*(1), 91 – 101.

Frank, O. (1980). Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference 4*(1), 45 – 50.

Frank, O. and T. Snijders (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics 10*, 53–53.

Frantz, T., M. Cataldo, and K. Carley (2009). Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory 15*(4), 303–328.

Gjoka, M., M. Kurant, C. T. Butts, and A. Markopoulou (2010). Walking in Facebook: A case study of unbiased sampling of OSNs. In *INFOCOM, 2010 Proceedings IEEE*, pp. 1–9.

Golder, S. A. and M. W. Macy (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science Magazine 333*, 1878–1881.

Golder, S. A. and S. Yardi (2010). Structural predictors of tie formation in Twitter: Transitivity and mutuality. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, Washington, DC, USA, pp. 88–95. IEEE Computer Society.

Goldstein, M. L., S. A. Morris, and G. G. Yen (2004). Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems 41*(2), 255–258.

Gonçalves, B., N. Perra, and A. Vespignani (2011, 08). Modeling users' activity on Twitter networks: Validation of Dunbar's Number. *PLoS one 6*.

Grannis, R. (2010). Six degrees of "Who cares?". *American Journal of Sociology 115*(4), 991–1017.

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

Grindrod, P. (2002). Range-dependent random graphs and their application to modeling large small-world Proteome datasets. *Physical Review E 66*, 066702.

Gruzd, A., S. Doiron, and P. Mai (2011). Is happiness contagious online? A case of Twitter and the 2010 Winter Olympics. In *2011 44th Hawaii International Conference on System Sciences (HICSS)*, pp. 1–9.

Guo, L., E. Tan, S. Chen, X. Zhang, and Y. E. Zhao (2009). Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, New York, NY, USA, pp. 369–378.

Han, J. D. J., D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology 23*, 839–944.

Hansen, N. (2005). The CMA evolution strategy: A tutorial. *Vu le 29*.

Hansen, N. and A. Ostermeier (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation 9*(2), 159–195.

Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, 174–199.

Heckathorn, D. D., S. Semaan, R. S. Broadhead, and J. J. Hughes (2002). Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18–25. *AIDS and Behavior 6*(1), 55–67.

Hill, A. L., D. G. Rand, M. A. Nowak, and N. A. Christakis (2010). Emotions as infectious diseases in a large social network: the SISa model. *Proceedings of the Royal Society B: Biological Sciences 277*(1701), 3827–3835.

Holme, P., B. J. Kim, C. N. Yoon, and S. K. Han (2002). Attack vulnerability of complex networks. *Physical Review E 65*(5), 056109.

Huberman, B. H., D. H. Romero, and F. Wu (2008). Social networks that matter: Twitter under the microscope. *CoRR abs/0812.1045*.

Hutto, C. J., S. Yardi, and E. Gilbert (2013). A longitudinal study of follow predictors on Twitter. In *CHI 2013 "Changing Perspectives," First ACM European Computing Research Congress*.

Jacomy, M., S. Heymann, T. Venturini, and M. Bastian (2012). Forceatlas2, a graph layout algorithm for handy network visualization. `http://www.medialab.sciences-po.fr/publications/Jacomy_Heymann_Venturini-Force_Atlas2.pdf`.

Java, A., X. Song, T. Finin, and B. Tseng (2009). Why we Twitter: An analysis of a microblogging community. In H. Zhang, M. Spiliopoulou, B. Mobasher, C. Giles, A. McCallum, O. Nasraoui, J. Srivastava, and J. Yen (Eds.), *Advances in Web Mining and Web Usage Analysis*, Volume 5439 of *Lecture Notes in Computer Science*, pp. 118–138. Springer Berlin / Heidelberg.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika 18*, 39–43.

Kim, E., S. Gilbert, M. J. Edwards, and E. Graeff (2009). Detecting sadness in 140 characters: Sentiment analysis and mourning Michael Jackson on Twitter. *Web Ecology 3*.

Kloumann, I. M., C. M. Danforth, K. D. Harris, C. A. Bliss, and P. S. Dodds (2012, 01). Positivity of the English language. *PLoS one 7*(1), e29484.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models* (1st ed.). New York, NY: Springer Publishing Company, Inc.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks 28*(3), 247–268.

Kossinets, G. and D. J. Watts (2006). Empirical analysis of an evolving social network. *Science 311*(5757), 88–90.

Kwak, H., C. Lee, H. Park, and S. Moon (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, New York, NY, USA, pp. 591–600.

Lakhina, A., J. Byers, M. Crovella, and P. Xie (2003, April). Sampling biases in IP topology measurements. In *Proceedings of IEEE Infocom*.

Lee, S. H., P.-J. Kim, and H. Jeong (2006). Statistical properties of sampled networks. *Physical Review E 73*(1), 016102.

Leroy, V., B. B. Cambazoglu, and F. Bonchi (2010). Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 393–402. ACM.

Leskovec, J. and C. Faloutsos (2006). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, New York, NY, USA, pp. 631–636. ACM.

Liben-Nowell, D. and J. Kleinberg (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology 58*(7), 1019–1031.

Lichtenwalter, R. N., J. T. Lussier, and N. V. Chawla (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 243–252. ACM.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Volume 1, pp. 296–304. San Francisco.

Lü, L. and T. Zhou (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications 390*(6), 1150–1170.

Lu, Z., B. Savas, W. Tang, and I. Dhillon (2010). Supervised link prediction using multiple sources. In *2010 IEEE 10th International Conference on Data Mining (ICDM)*, pp. 923–928. IEEE.

Lusseau, D., K. Schneider, O. Boisseau, P. Haases, E. Slooten, and S. Dawson (2003). The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology 54*, 396–405.

Lyons, R. (2011). The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy 2*(1), 1–26.

Martin, S., R. D. Carr, and J.-L. Faulon (2006). Random removal of edges from scale free graphs. *Physica A: Statistical Mechanics and its Applications 371*(2), 870 – 876.

Miller, G. (2011). Social scientists wade into the tweet stream. *Science Magazine 333*, 1814–1815.

Mitchell, L., M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one 8*(5), e64417.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley (2013). Is the sample good enough? Comparing data from Twitters streaming API with Twitters firehose. *Proceedings of ICWSM*.

Newman, M. E. J. (2001a, Jul). Clustering and preferential attachment in growing networks. *Physical Review E 64*, 025102.

Newman, M. E. J. (2001b). The structure of scientific collaboration networks. *Proceedings of the National Academy 98*, 404–409.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters 89*, 208701.

Noel, H., W. Galuba, and B. Nyhan (2011). The unfriending problem: The consequences of homophily in friendship retention for causal estimates of social influence. *Social Networks 33*, 211–218.

Papacharissi, Z. (2009). The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld. *New Media and Society 11*(1-2), 199–220.

Platig, J., M. Girvan, and E. Ott (2013). Robustness of network measures to link errors. *Bulletin of the American Physical Society 58*.

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science 27*(5), 292–306.

Rapoport, A. (1963). Mathematical models of social interaction. *Handbook of Mathematical Psychology 2*, 493–579.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási (2002). Hierarchical organization of modularity in metabolic networks. *Science 297*(5586), 1551–1555.

Romero, D. M., W. Galuba, S. Asur, and B. A. Huberman (2010). Influence and passivity in social media. *CoRR abs/1008.1253*.

Romero, D. M. and J. Kleinberg (2010). The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pp. 138–145.

Romero, D. M., B. Meeder, and J. Kleinberg (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of World Wide Web Conference*.

Romero, D. M., C. Tan, and J. Ugander (2013). On the interplay between social and topical structure. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM.

Rosenquist, J. N., J. Murabito, J. H. Fowler, and N. A. Christakis (2010). The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine 152*(7), 426–433.

Rowe, M., M. Stankovic, and H. Alani (2012). Who will follow whom? Exploiting semantics for link prediction in attention-information networks. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC'12, pp. 476–491.

Sadikov, E., M. Medina, J. Leskovec, and H. Garcia-Molina (2011). Correcting for missing data in information cascades. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, New York, NY, USA, pp. 55–64. ACM.

Salton, G. and M. J. McGill (1986). In *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Semaan, S., J. Lauby, and J. Liebman (2002). Street and network sampling in evaluation studies of hiv risk-reduction interventions. *AIDs Review 4*(4), 213–23.

Shalizi, C. R. and A. C. Thomas (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research 40*(2), 211–239.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika 42*(3/4), 425–440.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. skr. 5*, 1–34.

Stumpf, M., P. Ingram, I. Nouvel, and C. Wiuf (2005). Statistical model selection methods applied to biological networks. *Transactions on Computational Systems Biology III*, 65–77.

Stumpf, M. P. H., T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences 105*(19), 6959–6964.

Stumpf, M. P. H. and C. Wiuf (2005). Sampling properties of random graphs: the degree distribution. *Physical Review E 72*(3), 036118.

Stumpf, M. P. H., C. Wiuf, and R. M. May (2005). Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America 102*(12), 4221–4224.

Suttorp, T., N. Hansen, and C. Igel (2009). Efficient covariance matrix update for variable metric evolution strategies. *Machine Learning 75*(2), 167–197.

Tan, C., L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li (2011). User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1397–1405. ACM.

Taylor, A. and D. J. Higham (2009, February). CONTEST: A controllable test matrix toolbox for MATLAB. *ACM Transactions on Mathematical Software 35*, 26:1–26:17.

Thelwall, M., K. Buckley, and G. Paltoglou (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology 62*(2), 406–418.

Twitter (2011). Twitter API Blog. *http://blog.twitter.com/2011/09/one-hundred-million-voices*.

Twitter (2013). S-1 registration with the United States Securities and Exchange Commission (SEC). *http://www.sec.gov*.

Ugander, J., L. Backstrom, C. Marlow, and J. Kleinberg (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences 109*(16), 5962–5966.

Viswanath, B., A. Mislove, M. Cha, and K. P. Gummadi (2009). On the evolution of user interaction in Facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN '09, New York, NY, USA, pp. 37–42.

Wang, D., D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabási (2011). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, New York, NY, USA, pp. 1100–1108.

Wang, J. and L. Rong (2013). Similarity index based on the information of neighbor nodes for link prediction of complex network. *Modern Physics Letters B 27*(06).

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

Watts, D. J. and P. S. Dodds (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research 34*(4), 441–458.

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

Weng, J., E.-P. Lim, J. Jiang, and Q. He (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, New York, NY, USA, pp. 261–270.

Weng, L., F. Menczer, and Y.-Y. Ahn (2013). Virality prediction and community structure in social networks. *Scientific Reports*.

White, J., E. Southgate, J. Thompson, and S. Brenner (1986). The structure of the nervous system of the nematode *C. Elegans*. *Philosophical Transactions of the Royal Society of London 314*, 1–340.

Wiuf, C. and M. P. H. Stumpf (2006). Binomial subsampling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 462*(2068), 1181–1195.

Woolley-Meza, O., D. Grady, C. Thiemann, J. P. Bagrow, and D. Brockmann (2013). Eyjafjallajökull and 9/11: The impact of large-scale disasters on worldwide mobility. *PloS one 8*(8), e69829.

Yang, Y., N. V. Chawla, Y. Sun, and J. Han (2012). Predicting links in multi-relational and heterogeneous networks. In *Proceedings of the 12th IEEE International Conference on Data Mining*, ICDM 12, pp. 755–764.

Yin, D., L. Hong, and B. D. Davison (2011). Structural link analysis and prediction in microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, New York, NY, USA, pp. 1163–1168. ACM.

Yin, Z., M. Gupta, T. Weninger, and J. Han (2010). LINKREC: a unified framework for link recommendation with user attributes and graph structure. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 1211–1212. ACM.

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213*, 21–87.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 452–473.

Zhou, T., L. Lü, and Y. Zhang (2009). Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems 71*(4), 623–630.